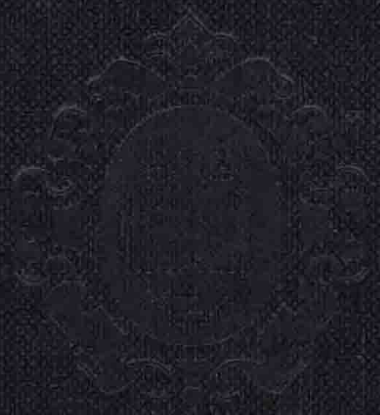
Roger Penrose
Collected Works

VOLUME 5

1990–1996

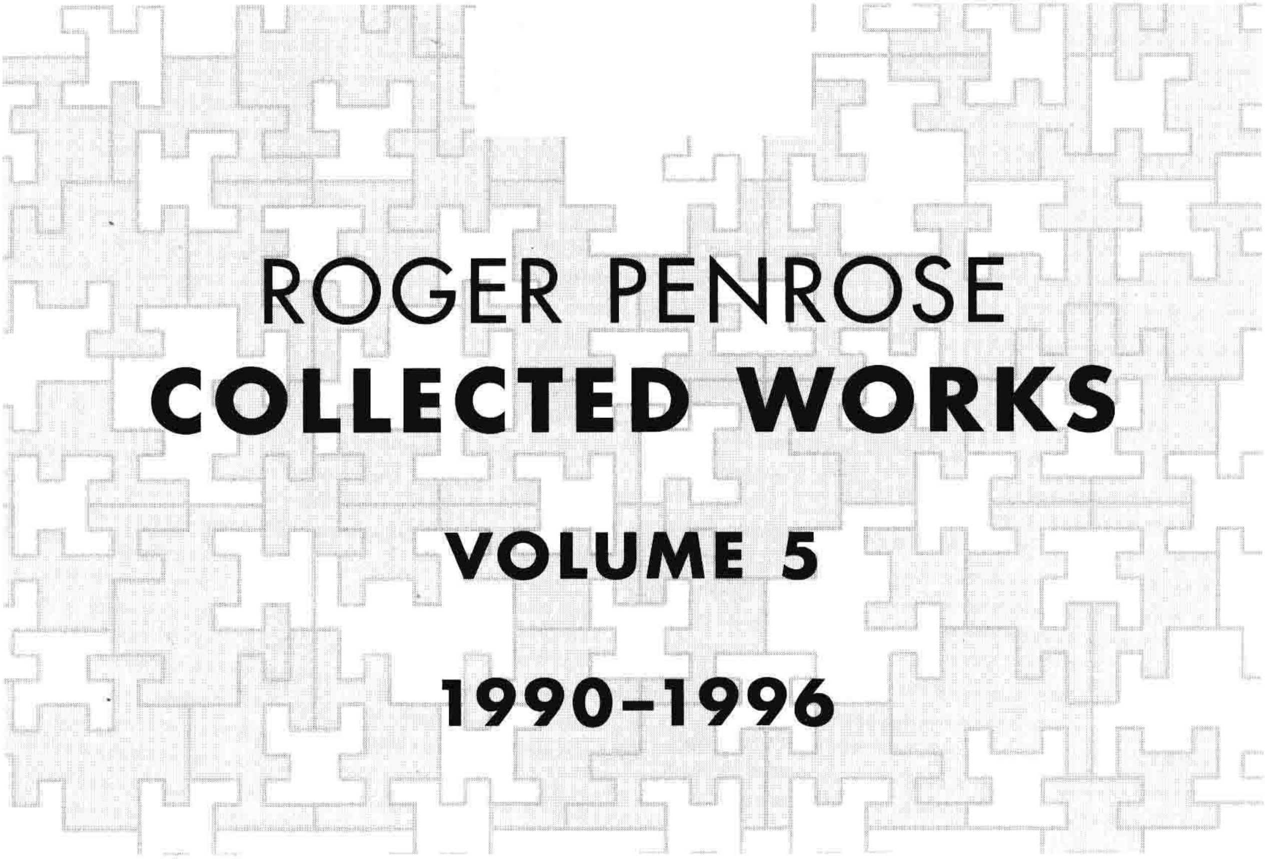


OXFORD SCIENCE PUBLICATIONS

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978 0 19 921940 7



ROGER PENROSE
COLLECTED WORKS

VOLUME 5

1990–1996

常州大学图书馆
藏书章

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Roger Penrose 2011
The moral rights of the authors have been asserted
Crown copyright material is reproduced under Class Licence
Number C01P0000148 with the permission of OPSI
and the Queen's Printer for Scotland
Database right Oxford University Press (maker)

First published 2011

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available
Library of Congress Control Number: 2010926298
Typeset by Newgen Imaging Systems Pvt Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire

ISBN 978–0–19–921940–7 (Vol. 5)
978–0–19–921944–5 (Set)

1 3 5 7 9 10 8 6 4 2

PREFACE

The topics addressed, in these *Collected Works*, vary over quite an extensive range, the boundaries between their areas being sometimes quite blurred. It would have been hard to classify all these articles and arrange them according to individual topics. Moreover, my own mode of working has been, on the whole, to keep most of my different scientific or mathematical interests running concurrently. Accordingly, it has seemed appropriate not to attempt to group the papers together into separate tidy categories. To do so would have provided a somewhat misleading picture of my overall mode of thinking. Instead, the ordering of articles that has been adopted is essentially chronological, at least with regard to the date of publication. The resulting somewhat higgledy-piggledy appearance that we find in the sometimes abrupt juxtaposition of different topics is, perhaps, fairly illustrative of my actual mode of working.

There are two main kinds of drive that underlie these works. On the one hand, some of the articles have arisen mainly from purely mathematical considerations, which can be ultimately aesthetic or just matters of simple curiosity. On the other hand, motivations for many others of the articles came primarily from a strong desire to discover more about the underlying principles governing the workings of the physical world. Yet, one of the deep mysteries about what has been discovered about these principles is the great breadth of intrinsically appealing mathematics that is found to underlie the actual operation of the physics of the universe, so these two drives are found to relate closely to one another. The reader may perhaps note the distinct geometrical flavour of much of my own mathematics, evidencing the strong appeal that geometry exerts on me, an appeal that is not infrequently intertwined with desire for algebraic elegance. Indeed, I have found it remarkable how much of the workings of our universe appear to depend upon sophisticated and powerful geometry and algebra.

My own very geometrical perspectives, as opposed to purely computational ones, are perhaps made more obvious in the later volumes of this series than in the early ones, but this appearance is deceptive. Very early in my career as a Cambridge graduate student, I had developed a graphical notation which enabled me to carry out many kinds of calculation in an entirely diagrammatic way. For a long time I was rather secretive of this procedure, and in my earlier writings would take care to remove all trace of these visual origins. I do not doubt that my predilections for such geometric or algebraically elegant forms of understanding have influenced the specific choices of direction in which my endeavours towards furthering our understandings of the physics of the world have taken.

The focus of these directions has become more clarified as time has moved on, and has narrowed down to such a degree that there has emerged a somewhat disturbing repetitiveness in many of the later articles. A good number of these provide rather non-technical accounts. An example of such repetitiveness occurs with my descriptions of gravitational collapse to a black hole. There is, however, a gradually developing outlook on this issue, leading to an emphasis on the fundamental contrast in character between the space-time singularities that inevitably appear within black holes and the space-time singularity that constitutes the big bang, this contrast being central to the existence of a second law of thermodynamics in our universe. The same kind of repetitiveness applies to my frequently repeated claim of a need for a fundamental change in the basic laws of quantum theory, where only towards the end do my suggestions become particularly specific. Likewise, many mildly differing accounts are provided outlining the formalism and the basic ideas that underlie twistor theory. The later ones lead up to various viewpoints on the 'googly' problem, which has constituted the main obstruction to the finding of a full understanding of general relativity in twistor terms. Despite such repetition, it is my hope that there is a sufficient variety of emphasis and of historical development in such articles for it to have been worth while to include them all here.

Several colleagues have collaborated with me in writing the works presented here, and I can mention but only a few. Foremost, I have collaborated long and particularly fruitfully with Ezra T. Newman, especially with regard to the asymptotics of general relativity, and also with Wolfgang Rindler in the development of the formalism of 2-spinor calculus. Other collaborators who have played important roles in helping to develop the mathematics of discrete physics and twistor theory have been Erwin Kronheimer, Michael Eastwood, Ronny Wells, and Denny Hill. Among my collaborators, providing very significant inputs, who were former students in my research group, are Andrew Hodges, Paul Tod, Lane Hughston, George Sparling, Richard Ward, Nick Woodhouse, and Lionel Mason, but there were many others. Although not playing roles as actual collaborators, fundamental influences in my research have come from Dennis Sciama and Michael Atiyah.

Roger Penrose
April 2010

BIOGRAPHY

Early years and influences

I was born in Colchester, Essex, UK on 8 August 1931. My father, Lionel Sharples Penrose, FRS, was a highly talented and intelligent man, with many interests. He, himself, had come from a Quaker family, his own father (James Doyle Penrose) having been a professional artist and his mother (Josephine Peckover) was from a well-to-do Quaker banking family. Lionel's primary professional interest was in the causes of hereditary mental disease. He was Galton Professor of Human Genetics at University College London from 1945 and, following retirement, Director of the Kennedy–Galton Institue, St Albans. He had a very considerable artistic talent and a keen interest in music, especially Mozart, but also Bach and other classical composers, dabbling a little in composition himself, particularly when there was a puzzle aspect in what was involved (where he once wrote a piece that could be played equally well upside down as right-way up). He liked making puzzles of various kinds in wood, and spent much of his later life designing wooden models illustrating self-reproduction. He was recognized as a fine composer of chess problems. He had a definite fondness for mathematics and was an innovative statistician. He loved explaining matters of science and mathematics. On one occasion, when I told him that my mathematics teacher at school had informed us that he would be explaining calculus the next day, Lionel immediately took me aside so as to be sure to be the first to teach me about the beauties of calculus. This made a great impression on me. Yet he was an extremely difficult person to communicate with on personal matters.

My mother, born Margaret Leathes, was a charming and caring woman with a very subtle mind. She had been Head Girl at Bedales school, and became medically qualified, but found she was not able to practice after her marriage. After Lionel died, in 1972, she married Max Newman FRS, a fascinating man, who had been Professor of Mathematics in Manchester and, as we later learned, a key figure at Bletchley Park during the war. Margaret's father, John Beresford Leathes, FRS, had been a distinguished professor of physiology, and her mother, a concert pianist, born Sara Mara Natanson from a Jewish family in Latvia, had disowned her family (then in St Petersberg), becoming completely secretive about her origins, and even about her maiden name. Margaret enjoyed geometry at school, but did not communicate much on this topic with me. My older brother Oliver was very precocious, and I learnt a lot of physics and mathematics from him as I was growing up. He later became a distinguished mathematical physicist (statistical mechanics) and FRS. My younger brother Jonathan excelled at games, particularly chess, and later became British Chess Champion a record 10 times, and then a top correspondence chess player. My much younger sister Shirley Victoria, now Shirley Hodgson, became a highly regarded Professor of Cancer Genetics.

This atmosphere of science, mathematics, art, and music, and of puzzles and games, had been an important part of my upbringing. During the war years, the family lived in London, Ontario, Canada, where Lionel became Director of Psychiatric Research at the Ontario Hospital. We returned to England in 1945, where I attended University College School, London. As an undergraduate, I studied Mathematics at University College London, from 1951 to 1954; then I went to St John's College Cambridge to do research in algebraic geometry, initially under William Hodge and subsequently John Todd. In my same year, under Hodge, was Micheal Atiyah, who was an inspirational, though somewhat intimidating colleague. Frank Adams and Christopher Zeemen (see Chapter 13) were contemporaries who also influenced me, and I became good friends with Hans Liebeck and Douglas Munn (see Chapter 3), who were close research-student colleagues at St John's.

Yet, the colleague who had the greatest influence on the development of my research during this period, and for many years later, was the cosmologist Dennis Sciama. Even before studying at Cambridge, I had a chance meeting with him during a visit to my brother Oliver in Cambridge in which I raised a question that had worried me in relation to the inspiring series of BBC radio talks given by Fred Hoyle in 1951 (The nature of the universe. A series of broadcast lectures). On the strength of this encounter (at which I used a space-time diagram with tilted light cones to make my point), Dennis had evidently formed the opinion that it would be worth while to interest me more in physics, even trying, later, to persuade me to switch my topic of research to cosmology. I was then too committed to my many pure-mathematical activities, but I learnt a great deal of physics from Dennis, and of the excitements involved in striving to uncover secrets of the universe. Another influential colleague was Felix Pirani, from whom I learnt much about developments in general relativity.

Pivotal were the superb lecture courses of Bondi, on general relativity, and Dirac, on quantum mechanics. On the pure-mathematical side, I recall particularly an elegant information-packed algebra course by Philip Hall. Lectures by Steen on mathematical logic turned out to influence me strongly towards my later viewpoint concerning the relation between computation and conscious understanding (e.g. Chapters 169, 184, 186).

Degrees

1952	BSc Mathematics (Spec) (1st Cl), University College, London
1957	PhD (Cantab), St John's College, Cambridge

Employment

1956	NRDC London (research on numerical analysis with computers); later consultant
1956–1957	Assistant Lecturer, Bedford College, London
1957–1960	Research Fellow, St John's College, Cambridge
1959–1961	NATO Research Fellow, Princeton, Syracuse, Cornell Universities
1961–1963	Research Associate, King's College, London
1963–1964	Visiting Associate Professor, University of Texas, Austin, Texas
1964–1966	Reader (Applied Mathematics), Birkbeck College, London
1966–1967	Visiting appointments: Yeshiva, Princeton, Cornell, Chicago Universities; lecturer Battelle Inst., Seattle
1967–1973	Professor (Applied Mathematics), Birkbeck College, London
1969–1970	Visiting professor, Yeshiva New York, Princeton, and Cornell Universities
1973–1998	Rouse Ball Professor (Mathematics), Oxford; 1998– Emeritus
1978	Miller Fellow, Univ. of California, Berkeley
1982–1987	Lovett Professor (Mathematics), Rice Univ., Houston, Texas, USA (Part-time)
1987–1993	Distinguished Professor (Physics, Mathematics), Syracuse Univ. USA (Part-time)
1993–	Francis and Helen Pentz Distinguished Professor (Physics, Mathematics; part-time)
1998–2001	Gresham Professor (Geometry), Gresham College, London

Honours Prizes include 1971 Heinemann Prize; 1972 Fellow of Royal Society; 1978 Plenary Lecture, International Congress of Mathematicians, Helsinki; 1985 Royal Medal (Royal Society);1988 Wolf Prize (physics, with Hawking); 1990 Albert Einstein Medal; 1994 knighted (services to science); 1994 Polish Academy of Sciences; 1998 US National Academy of Sciences; 2000 Order of Merit; 2004 de Morgan Medal (London Mathematical Society); 2001 Royal Irish Academy (Science); 2008 Copley Medal (Royal Society).

Marriages

1958	Joan Isabel Wedge; divorced 1980
1988	Vanessa Dee Thomas

Children

1963	Christopher Shaun
1964	Toby Nicholas
1966	Eric Alexander
2000	Maxwell Sebastian

Honorary degrees

1992	University of New Brunswick, Canada
1993	University of Surrey
1994	University of Bath
1995	University of London
1996	University of Glasgow
1996	University of Essex
1997	University of St Andrews
1998	University of Santiniketan, India
1998	Open University
2002	University of Southampton
2004	University of Waterloo, Canada
2005	University of Leuven, Belgium
2005	Warsaw University
2005	University of Athens
2006	University of York

Books

1972	Techniques of Differential Topology in Relativity (NSF and SIAM)
1984 and 1986	Spinors and Space-Time (Volumes 1 and 2; Cambridge University Press; with Rindler)
1989	The Emperor's New Mind (Oxford University Press 1989)
1994	Shadows of the Mind (Oxford University Press)
1996	The Nature of Space and Time (Princeton University Press; with Hawking)
1997	The Large, the Small and the Human Mind (Cambridge University Press)
2004	The Road to Reality (Jonathan Cape)

INTRODUCTION

Although my 1989 book *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, Oxford) had appeared shortly before the period covered by this Volume, numerous issues arose in relation to various people's responses to that work. The kind of vitriolic commentary that I sometimes encountered was something quite outside my earlier experiences, those usually having been within the more rationally-minded mathematical community, with its respect for reasoned argument. In Chapter 195, I provide an extended summary of the contents of my book, for the journal *Behavioral and Brain* Sciences, in relation to which they had organized peer review commentaries on this book from 37 different people (by no means all respectful). My responses to them are to be found in Chapter 194, and there were then five further critical commentaries from people taking issue with my responses, which I replied to in Chapter 225. Responses to other viewpoints of relevance are to be found in Chapters 206 and 209. One extremely valuable reaction to this book was a communication from the American anaesthesiologist Stuart Hameroff, who acquainted me with the presence and the importance of microtubules in relation the ideas that I had been proposing in *The Emperor's New Mind.* I soon realized that these structures might indeed supply the missing ingredient for the proposals that I had been putting forward, and this correspondence from Hameroff began a collaboration between us that resulted in a moderately specific scheme whereby large-scale quantum coherence might exist within the conscious brain, and actual conscious experience would be associated with an objective reduction of the quantum state (Chapters 246 and 247).

In twistor theory, I put forward an idea that a twistor might usefully be regarded as a source (or 'charge') for a massless field of spin 3/2, and since the consistency condition for the existence of such massless fields is Ricci-flatness, a pertinent definition for the twistor space for a Ricci-flat space-time might actually be as the space of such possible sources. A large number of articles (largely in *Twistor Newsletter*) were written during this period, in attempts to develop this seemingly promising idea, but it never seemed possible to resolve all the difficulties that continued to confront it. Towards the end of the period covered by this Volume, I began to develop a quite different approach to incorporating full general relativity (storing the googly information in appropriate differential forms) which would ultimately lead to a much more promising-looking development.

A significant development of a different kind was a change of heart that I had, while in Argentina trying to prepare my Plenary talk for the International Conference on General Relativity and Gravitation (while still suffering from the after-effects of a nasty bout of toxoplasmosis that I had contracted in France a couple of months earlier). I performed a calculation that led me to change the proposal that I had previously been promoting concerning the idea of a spontaneous reduction of quantum states when the gravitational effects of mass displacements become significant (Chapters 215, 218, and 252). Some years earlier (though motivated by quite different considerations) Lajos Diósi had had a very closely related idea, and this is sometimes known as the Diósi–Penrose proposal.

ACKNOWLEDGEMENTS

Oxford University Press has attempted to trace the copyright holder of all the papers included in these volumes and apologizes to copyright holders if permission to publish in this form has not been obtained.

The author and publishers thank the following individuals for their permissions:

- *Acta Physics Polonica B* **280**
- Adam Hilger Ltd **147**
- American Chemical Society **206**
- American Institute of Physics
  - *American Journal of Physics* **19, 24**
  - *Journal of Maths and Physics* **15, 26, 27, 29, 39, 41, 58, 84**
  - *Errata* **15**
- American Mathematical Society **30, 180**
  - *Bulletin of the American Mathematical Society* **144**
  - *Contemporary Mathematics* **179**
- American Physical Society
  - *Physical Review Letters* **16, 20, 23, 302**
  - *Reviews of Modern Physics* **21**
- *Annals of Human Genetics* **100**
- *Asian Journal of Mathematics* **281**
- Astronomische Gesellschaft **295**

- Benjamin **36**
- Bibliopolis **175**
- Blackwell **80, 166, 186, 215**
  - *Annals of the New York Academy of Sciences* **60, 190, 294**
- British Broadcasting Company **59**
- *British Journal of Psychology* **9**

Cambridge Entomological Club **243**

Cambridge University Press **6, 7, 46, 104, 138, 157, 172, 173, 176, 195, 218, 220, 227, 230, 240, 290, 303, 304**

  Behavioural and Brain Sciences **193, 194, 224**

  Proceedings of the Cambridge Philosophical Society **3, 4, 5, 32, 42, 50, 126**

Chapman and Hall/CRC Press **107, 124, 125, 127, 128, 130, 136, 139, 142, 143, 146, 149, 150, 154, 155, 156, 160, 161, 164, 170, 171, 178, 189, 191, 198, 201, 203, 209, 229, 234, 293**

Connected Globe **257**

Consigle Nazionale Della Richerche **145**

Cornell University Press **33**

Éditions du Centre National de la Recherche Scientifique **14**

Elsevier **12, 45, 52, 68, 85, 86, 92, 93, 123, 151, 167, 185, 187, 226, 279**

Escher, M.C. images: The M.C. Escher Company B.V. **218, 249, 306**

Eureka **1, 2**

Folio Society **284**

Freeman **56**

Granta **297**

Gresham College **277**

Halsted Press **148**

Hanser **66, 67**

Imperial College Press **264**

Imprint Academic **235, 245**

Indian Academy of Sciences **244**

Indiana University Press **25**

Institute of Electrical and Electronics Engineers **196**

Institute of Physics **217**

  Classical and Quantum Gravity **253, 254, 274, 282**

  European Journal of Physics **258**

  Journal of Physics A: Mathematical and General **95**

International Astronomical Union

  International Astronomical Union Symposium Proceedings Series No. 13, Cosmology **72**

International Atomic Energy Agency
  Contemporary Physics **38**
International Centre for Theoretical Physics **53, 68**
International Mathematical Union **286**
International University Bremen **305**
Inter-University Centre for Astronomy and Physics **265**
The Italian Astronomical Society
  Memorie della Società Astronomica Italiana **200**

Longman **78, 79, 108, 116, 117, 134, 135, 141, 177, 188**

The Massachusetts Institute of Technology Press **228, 246**

National Institute of Sciences of Ukraine **48**
Nature **8, 44, 49**
Nature (Physical Science) **47**
New Scientist **10, 101**
New York Academy of Sciences **294**
Oxford University Press **55, 69, 121, 122, 165, 178, 184, 197, 202, 208, 216, 237, 270, 272, 276, 291, 299**

Pearson **109**
Penguin **273**
Penrose's Unpublished Papers **31**
Pergamon Press **17**
Pitman **73, 75, 76, 77, 82, 87, 89, 90, 94, 102, 103, 110**
Pontificial University **296**
Princeton University Press **28, 268**
  Annals of Math **13, 131**
Proceedings of the International Congress of Mathematicians **115**

Rivista del Nuovo Cimento **40**
Routledge **169**
Royal Astronomical Society **183**
The Royal Society
  Proceedings of the Royal Society Series A **22, 35, 97, 105, 106, 132**
  Philosophical Transactions of the Royal Society of London. Series A **275, 283**

- The Sciences **99**
- *Scientific American* **51**
- Simon and Schuster **242**
- Springer **61, 63, 65, 83, 91, 98, 111, 133, 168, 204, 207, 231, 255, 263, 278, 292, 298, 300**
  - *Artificial Intelligence Review* **205**
  - *Communications in Mathematical Physics* **43, 54, 129**
  - *Cosmology and Gravitation* **120**
  - *Foundations of Physics* **140**
  - *General Relativity and Gravitation* **70, 71, 113, 114, 239, 251**
  - *Higher-Order and Symbolic Computation* **287**
  - *International Journal of Theoretical Physics* **37, 56**
  - *Journal of Statistical Physics* **233**
  - *Lecture Notes in Physics* **112, 152**
  - *The Mathematical Intelligencer* **285**
- Taylor and Francis **18**
  - *International Studies in the Philosophy of Science* **256**
  - *Surveys in High Energy Physics* **119**
- Trinity College, Dublin **219**

- World Scientific **162, 163, 225, 236, 241, 252, 289, 301**

- Universal Academy Press **266**
- University of Chicago Press **96, 269**
- University of Chittagong **158**
- University College London **267**

- The University of Oxford (Mathematics Department)
  - *Twistor Newsletter* **73, 74, 75, 76, 77, 78, 79, 81, 82, 88, 89, 90, 94, 102, 103, 107, 108, 116, 117, 118, 124, 125, 127, 128, 130, 134, 135, 136, 137, 139, 141, 142, 143, 146, 149, 150, 153, 154, 155, 156, 157, 160, 161, 164, 170, 171, 174, 181, 182, 188, 189, 191, 192, 198, 199, 201, 203, 209, 210, 211, 212, 213, 214, 221, 222, 223, 229, 232, 238, 248, 290, 250, 259, 260, 261, 262, 271, 288**

- University of Utah Press **247**

CONTENTS
VOLUME 5

Introduction *xxv*

1990s

192 Difficulties with inflationary cosmology 1
193 Light rays near i^0: A new mass-positivity theorem 19
194 Extended Regge trajectories: 12 years after 27
195 Précis of *The Emperor's New Mind: Concerning computers, minds and the laws of physics* 29
196 Author's response – the nonalgorithmic mind 43
197 Twistor theory after 25 years – its physical status and prospects 59
198 Twistors, particles, strings and links 89
199 Mass positivity from focussing and the structure of space-like infinity 101
200 Twistor theory for vacuum space-time: A new approach 107
201 What does the big bang tell us about quantum gravity? 113
202 Twistors as spin 3/2 charges 123
203 Twistors as charges for spin 3/2 in vacuum 133
204 The mass of the classical vacuum 143
205 Twistors as spin 3/2 charges continued: SL(3,C) bundles 151
206 Response to Tony Dodd's 'Godel, Penrose and the possibility of AI' 159
207 The biggest enigma 161
208 $\mathcal{H}$-space and twistors 167
209 'Penrose and mathematical ability' 189
210 The most general {2,2} self-dual vacuum: A googly approach 193
211 On impulsive gravitational waves 199
212 Kinking and causality 205
213 Puzzle page 213
214 Solution to puzzle in TN34 215
215 Gravitational collapse of the wave function: New thoughts 221
216 Setting the scene: The claim and the issues 225
217 The modern physicist's view of nature 259

218 Gravity and quantum mechanics 311
219 Foreword to *Mr Tompkins in paperback* 323
220 Mathematics, physics, computers and conscious understanding 327
221 An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves 339
222 Orthogonality of general spin states 349
223 A simple observation concerning {2 2} vacuums 355
224 The orthographic image of a regular tetrahedron 357
225 Author's response – An Emperor still without mind 361
226 Square-root voting 369
227 On Bell non-locality without probabilities: More curious geometry 375
228 Quantum non-locality and complex reality 401
229 On the cohomology of impossible figures 415
230 Spin 3/2 fields and local twistors 419
231 Mathematical intelligence 427
232 Is conscious awareness consistent with space-time descriptions? 459
233 A twistor-topological approach to the Einstein equations 475
234 On the second law of thermodynamics 485
235 Twistors and the Einstein equations 491
236 What 'gaps'? Reply to Grush and Churchland 505
237 Non-locality and objectivity in quantum state reduction 521
238 Must mathematical physics be reductionist? 531
239 Concerning space-time points for spin 3/2 twistor spaces 547
240 In {2 2} vacuums, the null-datum on null rays is exactly r^{-5} 553
241 Why new physics is needed to understand the mind 557
242 The role of gravity in quantum state reduction 573
243 Consciousness involves noncomputable ingredients 593
244 Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind* 603
245 Chandrasekhar, black holes, and singularities 641
246 Conscious events as orchestrated space-time selections 661
247 Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness? 681
248 Space-time and cosmology 711

249 New ideas for the googly graviton 757
250 Incidence between complex null rays 771
251 Two puzzles 777
252 On gravity's role in quantum state reduction 779
253 Twistor theory, the Einstein equations, and quantum mechanics 801

DETAILED CONTENTS
VOLUMES 1–6

VOLUME 1

1950s

1 Morley's trisector theorem 1
2 'Construction problem' and 'Word problem' 3
3 A note on inverse semigroups 5
4 A generalized inverse for matrices 11
5 On best approximate solutions of linear matrix equations 21
6 Tensor methods in algebraic geometry 25
7 Geometrical algebras: A new approach to invariant theory 251
8 A self-reproducing analogue 351
9 Impossible objects: A special type of visual illusion 355
10 Puzzles for Christmas 359
11 The apparent shape of a relativistically moving sphere 365

1960s

12 A spinor approach to general relativity 369
13 Imbedding of manifolds in Euclidean space 403
14 General relativity in spinor form 415
15 An approach to gravitational radiation by a method of spin coefficients 421
16 Asymptotic properties of fields and space-times 437
17 The light cone at infinity 441
18 Conformal treatment of infinity 447
19 Energy conservation as the basis of relativistic mechanics 469
20 Gravitational collapse and space-time singularities 475
21 A remarkable property of plane waves in general relativity 479
22 Zero rest-mass fields including gravitation: Asymptotic behaviour 487
23 10 exact gravitationally conserved quantities 533
24 Energy conservation as the basis of relativistic mechanics II 537
25 General-relativistic energy flux and elementary optics 541
26 Note on the Bondi-Metzner-Sachs group 559
27 Some new gravitationally conserved quantities 569
28 An analysis of the structure of space-time 579
29 Twistor algebra 731
30 Conserved quantities and conformal structure in general relativity 755
31 Theory of quantized directions 769
32 On the structure of causal spaces 801

VOLUME 2

1960s

33 Cosmological boundary conditions for zero rest-mass fields 1
34 On gravitational collapse and cosmology 15
35 New conservation laws for zero rest-mass fields in asymptotically flat space-time 29
36 Structure of space-time 61
37 Twistor quantization and curved space-time 177
38 On gravitational collapse 217

39 Solutions of the zero-rest-mass equations 231
40 Gravitational collapse: The role of general relativity 235
41 Conserved quantities in the Einstein–Maxwell theory 261

1970s

42 The singularities of gravitational collapse and cosmology 267
43 On quadratic first integrals of the geodesic equations for type {22} spacetimes 289
44 Scattering of two impulsive gravitational plane waves 301
45 Applications of negative dimensional tensors 307
46 Angular momentum: An approach to combinatorial space-time 333
47 Extraction of rotational energy from a black hole 367
48 Spinor classification of energy tensors 375
49 Black holes and gravitational theory 389
50 Ideal points in space-time 395
51 Black holes 419
52 Twistor theory: An approach to the quantisation of fields and space-time 429
53 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants' 505
54 On a quadratic first integral for the charged particle orbits in the charged Kerr solution 513
55 The geometry of impulsive gravitational waves 521
56 On the nature of quantum geometry 537
57 Internal instability in a Reissner–Nordström black hole 583
58 A space-time calculus based on pairs of null directions 599
59 Black holes 609
60 Naked singularities 635
61 Relativistic symmetry groups 647
62 Gravitational collapse 707
63 Gravitational collapse 729
64 The role of aesthetics in pure and applied mathematical research 741
65 Singularities in cosmology 749
66 Twistors and particles: An outline 761
67 The non-linear graviton 779
68 Classical gravity: A status report 791
69 Twistor theory: Its aims and achievements 795

VOLUME 3

1970s

70 Nonlinear gravitons and curved twistor theory 1
71 The nonlinear graviton 25
72 The space-time singularities of cosmology and in black holes 33
73 The twistor quadrille 45
74 The snail contour 53
75 The universal bracket factor 55
76 Some elementary twistor integrals with boundary 59
77 Twistor functions and sheaf cohomology 63
78 How to grow (or at least half grow) new twistors from old 77
79 Some remarks on the two twistor description of electromagnetic and linearized gravitational potentials 81
80 The cosmic censorship hypothesis 83
81 A finite covering for PT^+ 87
82 The back-handed photon (or, to a cricketer, the googly photon!) 89
83 Any space-time has a plane wave as a limit 99
84 The Kähler structure of asymptotic twistor space 105
85 The twistor programme 113
86 Space-time singularities 127
87 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel 137
88 Remarks concerning the ambidextrous photon 145
89 Massless fields and sheaf cohomology 149
90 Note on the n-twistor internal symmetry group 159
91 Aspects of general relativity 163
92 Is nature complex? 203
93 Gravitational collapse: A review 213
94 Local H^1's and propagation 231

95 Extended Regge trajectories 237
96 Singularities of spacetime 243
97 The metric and curvature properties of $\mathcal{H}$-space 271
98 The geometry of the universe 297
99 Einstein's vision and the mathematics of the natural world 341
100 The topology of ridge systems 347
101 Twisting round space-time 359
102 A googly graviton? 365
103 Twistors as helicity raising operators 375
104 Singularities and time-asymmetry 381
105 General introduction 441
106 Global general relativity 445
107 Remarks on the Sparling and Eguchi-Hanson (googly?) gravitons 449
108 The anti-self-dual Coulomb field's non-Hausdorff twistor space 459
109 On the twistor descriptions of massless fields 461
110 Combinatorial quantum theory and quantized directions 499
111 Pentaplexity: A class of nonperiodic tilings of the plane 507

1980s

112 Recent advances in global general relativity: A brief survey 515
113 Twistors for flat and curved space-time 527
114 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity 575
115 The complex geometry of the natural world 617
116 On the evaluation of twistor cohomology classes 625
117 Three channels for the box diagram 629
118 Puzzle: A sequence for twistorians 637
119 A brief introduction to twistors 639
120 A brief outline of twistor theory 663

VOLUME 4

1980s

121 Time-asymmetry and quantum gravity 1
122 Some remarks on twistors and curved-space quantization 31
123 On Schwarzschild causality—a problem for "Lorentz covariant" general relativity 47
124 Concerning a Fourier contour integral 61
125 A new angle on the googly graviton 65
126 A quadratic mapping with invariant cubic curve 73
127 Twistors for cosmological models 91
128 Physical left-right symmetry and googlies 97
129 Cohomology and massless fields 105
130 A theory of 2-surface ('superficial') twistors 153
131 Some unsolved problems in classical general relativity 161
132 Quasi-local mass and angular momentum in general relativity 201
133 Time-asymmetry, cosmological uniformity and space-time singularities 213
134 Real-space topology for advanced/retarded twistor functions 217
135 Twistor blisters and relative cohomology 221
136 On the geometry of googly maps 225
137 A puzzle for twistorians 229
138 Some remarks on gravity and quantum mechanics 231
139 General-relativistic kinematics? 241
140 Spinors and torsion in general relativity 247
141 Non-locality in physics: A new mathematical viewpoint 263
142 On closed-set coverings for PT$^+$—A correction 269
143 Remarks on curved-space, twistor theory and googlies 271
144 A note on Sparling's 3-form 277
145 Physical space-time and nonrealizable CR-structures 281
146 Quasi-local mass and angular momentum 305
147 Relative cohomology, googlies and deformations of $\mathbb{I}$ 309
148 Donaldson's moduli space: A 'model' for quantum gravity? 313
149 Mass in general relativity 319
150 'New improved' quasi-local mass, and the Schwarzschild solution 331
151 On Bryant's condition for holomorphic curves in CR space 339

152 Integrals for general-relativistic sources: A development from Maxwell's electromagnetic theory 343
153 Mass and angular momentum at the quasi-local level in general relativity 377
154 Objective state-vector reduction? 387
155 Embedding 2-surfaces in CM 393
156 A suggested further modification to the quasi-local formula 397
157 Higher-dimensional two-surface twistors 399
158 Numerical relativity by power series 403
159 Elements of twistor theory 409
160 Energy and its definition in general relativity 429
161 Two-surface twistors and hypersurface twistors 441
162 Exponentiating a relative H^2 445
163 Gravitational field mass 449
164 Gravity and quantum theory 467
165 Twistors and the time-irreversability of state-vector reduction 473
166 Gravity and state vector reduction 477
167 Big bangs, black holes and 'time's arrow' 497
168 Escher and the visual representation of mathematical ideas 525
169 Hermann Weyl, space-time and conformal geometry 541
170 Quantum physics and conscious thought 573
171 Local twistor transport at $\mathscr{I}^+$: An approach to the googly 601
172 An approach to a coordinate-free calculus at $\mathbb{I}$ 609
173 Twistors in general relativity 613
174 Newton, quantum theory and reality 633
175 On Qadir's intertwining operators 667
176 On the origins of twistor theory 671
177 New directions in quantum gravity 693
178 Pretzel twistor spaces 703
179 Twistors and state-vector reduction 713
180 Aspects of quasi-local angular momentum 717
181 Fundamental asymmetry in physical laws 727
182 Topological QFT and twistors: Holomorphic linking 741
183 Holomorphic linking: Postscript 745
184 Singularities and big-bang cosmology 747
185 On the physics and mathematics of thought 751
186 A twistor conformal field theory for four space-time dimensions 785
187 Minds, machines and mathematics 791
188 Tilings and quasi-crystals: A non-local growth problem? 811
189 Non-Hausdorff Riemann surfaces and complex dynamical systems 839
190 Using spinors in general relativity 845
191 A twistorial approach to the full vacuum equations 851

VOLUME 5

1990s

192 Difficulties with inflationary cosmology 1
193 Light rays near i^0: A new mass-positivity theorem 19
194 Extended Regge trajectories: 12 years after 27
195 Précis of *The Emperor's New Mind: Concerning computers, minds and the laws of physics* 29
196 Author's response – the nonalgorithmic mind 43
197 Twistor theory after 25 years – its physical status and prospects 59
198 Twistors, particles, strings and links 89
199 Mass positivity from focussing and the structure of space-like infinity 101
200 Twistor theory for vacuum space-time: A new approach 107
201 What does the big bang tell us about quantum gravity? 113
202 Twistors as spin 3/2 charges 123
203 Twistors as charges for spin 3/2 in vacuum 133
204 The mass of the classical vacuum 143
205 Twistors as spin 3/2 charges continued: SL(3,C) bundles 151
206 Response to Tony Dodd's 'Godel, Penrose and the possibility of AI' 159
207 The biggest enigma 161
208 $\mathcal{H}$-space and twistors 167
209 'Penrose and mathematical ability' 189

210 The most general {2,2} self-dual vacuum: A googly approach 193
211 On impulsive gravitational waves 199
212 Kinking and causality 205
213 Puzzle page 213
214 Solution to puzzle in TN34 215
215 Gravitational collapse of the wave function: New thoughts 221
216 Setting the scene: The claim and the issues 225
217 The modern physicist's view of nature 259
218 Gravity and quantum mechanics 311
219 Foreword to *Mr Tompkins in paperback* 323
220 Mathematics, physics, computers and conscious understanding 327
221 An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves 339
222 Orthogonality of general spin states 349
223 A simple observation concerning {2 2} vacuums 355
224 The orthographic image of a regular tetrahedron 357
225 Author's response – An Emperor still without mind 361
226 Square-root voting 369
227 On Bell non-locality without probabilities: More curious geometry 375
228 Quantum non-locality and complex reality 401
229 On the cohomology of impossible figures 415
230 Spin 3/2 fields and local twistors 419
231 Mathematical intelligence 427
232 Is conscious awareness consistent with space-time descriptions? 459
233 A twistor-topological approach to the Einstein equations 475
234 On the second law of thermodynamics 485
235 Twistors and the Einstein equations 491
236 What 'gaps'? Reply to Grush and Churchland 505
237 Non-locality and objectivity in quantum state reduction 521
238 Must mathematical physics be reductionist? 531
239 Concerning space-time points for spin 3/2 twistor spaces 547
240 In {2 2} vacuums, the null-datum on null rays is exactly r^{-5} 553
241 Why new physics is needed to understand the mind 557
242 The role of gravity in quantum state reduction 573
243 Consciousness involves noncomputable ingredients 593
244 Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind* 603
245 Chandrasekhar, black holes, and singularities 641
246 Conscious events as orchestrated space-time selections 661
247 Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness? 681
248 Space-time and cosmology 711
249 New ideas for the googly graviton 757
250 Incidence between complex null rays 771
251 Two puzzles 777
252 On gravity's role in quantum state reduction 779
253 Twistor theory, the Einstein equations, and quantum mechanics 801

VOLUME 6

1990s

254 Andrzej Mariusz Trautman 1
255 Twistor geometry of light rays 11
256 Remarks on tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set 37
257 On understanding understanding 69
258 Mathematics: Pattern, precision and profundity 85
259 The mathematics of the electron's spin 109
260 Five things you can do with a surface in PT 115
261 Solutions to puzzles in TN41 121
262 Corrections to missprints in solutions to TN41 puzzles 125
263 Googly maps as forms 127
264 The need for a non-computational extension of quantum action in the brain 133
265 Chandrasekhar and the end of time 151
266 Science and the mind 161

267 Quantum state-reduction as a gravitational phenomenon; theory and experiment 165
268 Lionel Penrose: Colleague and father 169
269 Foreword to *Einstein's miraculous year: Five papers that changed the face of physics* 175
270 The question of cosmic censorship 185
271 Afterword to *The geometric universe: Science, geometry, and the work of Roger Penrose* 203
272 The twistor space for a strongly asymptotically flat vacuum 213
273 The complexity of our singular universe 223
274 Can a computer understand? 251
275 Spherically-symmetric solutions of the Schrödinger–Newton equations 279
276 Quantum computation, entanglement and state reduction 291
277 Foreword to *Ludwig Boltzmann: The man who trusted atoms* 305
278 Mathematics in the arts 309
279 Some remarks on twistor theory 323
280 The central programme of twistor theory 339
281 Lecture in honour of Leopold Infeld (extended outline only): Spinors in general relativity 371
282 On Ricci-flat twistor theory 381
283 Twistor theory and the Einstein vacuum 389
284 Introduction: Science at the turn of the millennium. 'Science into the next millennium: Young scientists give their visions of the future, part I.' 409

2000s

285 Introduction. *What is life?* 417
286 Drawing with complex numbers 433
287 Mathematical physics of the 20th and 21st centuries 441
288 Reminiscences of Christopher Strachey 459
289 On extracting the googly information 463
290 Wavefunction collapse as a real gravitational effect 489
291 On Bell non-locality without probabilities: Some curious geometry 509
292 The heritage of Pythagoras: Nineteen to the dozen 537
293 Twistors and general relativity 551
294 Towards a twistor description of general space-times: Introductory comments 579
295 Consciousness, the brain, and spacetime geometry: An addendum 597
296 The Schwarzschild singularity 605
297 The physics of the small and large: What is the bridge between them? 621
298 The rediscovery of gravity. The Einstein equation of general relativity 629
299 John Bell, state reduction, and quanglement 663
300 Consciousness, computation, and the Chinese room 677
301 Twistor geometry of conformal infinity 703
302 Gravitational collapse of the wavefunction: An experimentally testable proposal 713
303 Towards quantum superpositions of a mirror 719
304 On the instability of extra space dimensions 725
305 The problem of spacetime singularities: Implications for quantum gravity? 743
306 The remarkable role of mathematics in the physical universe 767

192

DIFFICULTIES WITH INFLATIONARY COSMOLOGY

Fourteenth Texas Symposium on Relativistic Astrophysics. Ed. Ervin J. Feyves, 249–64. *Annals of the New York Academy of Sciences,* Vol. 571, 1990

This is my earliest expression of disquiet concerning inflationary cosmology, where I point out that most of the early claims that inflation provides a satisfactory explanation for certain basic puzzles of cosmology are unjustified, and that the real puzzle is a much more basic one concerning the origin and form of the second law of thermodynamics. Although, since then, I have not changed my position concerning any of my own arguments, and I still seriously doubt the reality of inflation as a true feature of the very early universe, I have to confess to being impressed by the fact that various later observations have conformed well to predictions made by inflation. If *not* due to inflation, these would need alternative explanations. My own standpoint is still in opposition to inflation (as of May 2010) and my alternative position is put forth in *Cycles of Time* (Bodley Head, 2010).

Difficulties with inflationary cosmology. *Fourteenth Texas Symposium on Relativistic Astrophysics.* Ed. Ervin J. Feyves, 249–64. *Annals of the New York Academy of Sciences,* Vol. 571, 1990

Difficulties with Inflationary Cosmology[a]

ROGER PENROSE

Mathematical Institute
Oxford University
Oxford OX1 3LB, United Kingdom
and
Department of Physics
Syracuse University
Syracuse, New York 13210

CAN INFLATION ACHIEVE ITS GOALS?

The idea of inflationary cosmology is an ingenious attempt to solve some of the major puzzles of cosmology, most notably the flatness problem, the homogeneity (horizon) problem, and the monopole problem (see references 1–5; cf. Börner[6] for a critical evaluation). To what extent can this idea really provide a solution to problems of this kind? I do not believe that the issue has been properly addressed. The homogeneity problem, in particular, is intimately connected with a largely unappreciated, but profound puzzle presented to us by the second law of thermodynamics. The mechanism of inflation does not, by itself, come to terms with this and consequently, as we shall see, comes nowhere close to providing us with an understanding of the large-scale homogeneity of the universe (cf. also reference 7).

I should first remark that the problems that inflation attempts to resolve are really of two quite distinct kinds: "external" and "internal". The monopole problem is an example of an internal (or "self-made") problem, by which I mean it is a problem thrown up by the particular type of theory (appropriate GUT model with symmetry breaking) that is itself invoked as a principal ingredient of the inflationary mechanism. The observational absence of a proliferation of monopoles could just as well be taken as evidence against the type of theory that predicts their presence as taken as evidence in favor of the inflationary implications that such a theory might have. Apart from this, I have no quarrel with the arguments that use inflation to resolve problems of this kind; it is a necessary self-consistency requirement for such a theory.

The flatness and homogeneity problems, on the other hand, are external problems in the sense that they are directly presented to us as facts about the universe, irrespective of what views we might happen to hold concerning the physical validity of particular models of particle physics (GUT, symmetry breaking, etc.; see, for example, reference 8 for a discussion of a possible fundamental alternative to the symmetry-breaking idea). I shall concentrate here on the homogeneity problem and its relation to the second law of thermodynamics. I have explained this relationship many times elsewhere (cf. references 8–11), so I do not think that it is necessary for me to spell it out in full detail here; only the essentials of the argument will be presented.

---

[a] This research was supported by NSF Grant No. PHY86-12424. The figures used herein are printed with permission from Oxford University Press (*The Emperor's New Mind* by Roger Penrose, 1989).

THE SECOND LAW OF THERMODYNAMICS

According to the second law, the entropy of any isolated system should be nondecreasing with time and, indeed, this entropy increase should apply also to the universe as a whole. There is a difficulty with this law in that the concept "entropy" is not always very well defined when referring to physical systems generally. However, it is a striking fact that the uncertainties of definition seem to be of little consequence in practice. For a general concept of entropy, we need a coarse-graining of phase-space into regions representing states of the system that are considered to be macroscopically indistinguishable; then, the entropy, S, is proportional to the logarithm of the phase-space volume, V, of that region:

$$S = k \log V,$$

where k is Boltzmann's constant. The uncertainty in the definition of S arises from an arbitrariness in our choice of criterion for what "macroscopically indistinguishable" is actually to mean. The reason that this seems to be of little consequence in practice is that the differences (ratios) between the various phase-space volumes that we are normally concerned with are absolutely enormous and totally swamp any differences that arise from "reasonable" changes in our criterion of distinguishability.

To understand the second law, we can think of the phase-space point representing our system as being located in some very tiny coarse-graining region, representing an initial state of low entropy, and then the point wanders around in the phase-space, finding larger and larger regions. Each successive region is so much larger than the previous one that there is, in effect, no chance that the point will find any one of the

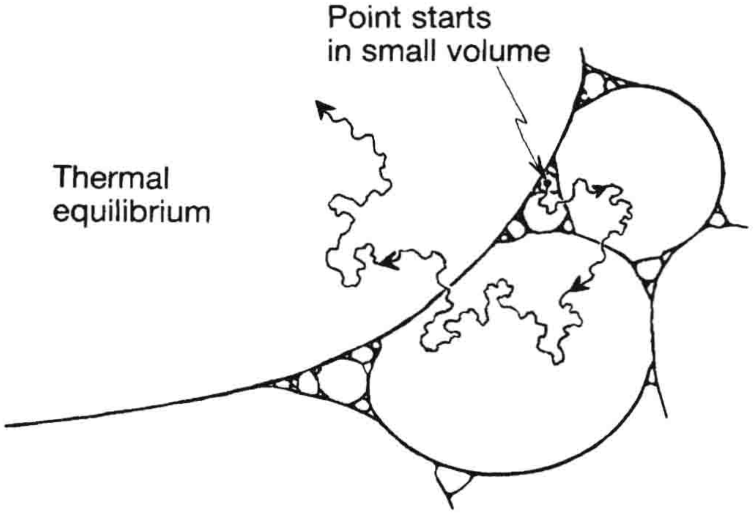


FIGURE 1. A coarse-grained phase-space; the second law of thermodynamics in action.

4 Difficulties with inflationary cosmology. *Fourteenth Texas Symposium on Relativistic Astrophysics*. Ed. Ervin J. Feyves, 249–64. *Annals of the New York Academy of Sciences*, Vol. 571, 1990

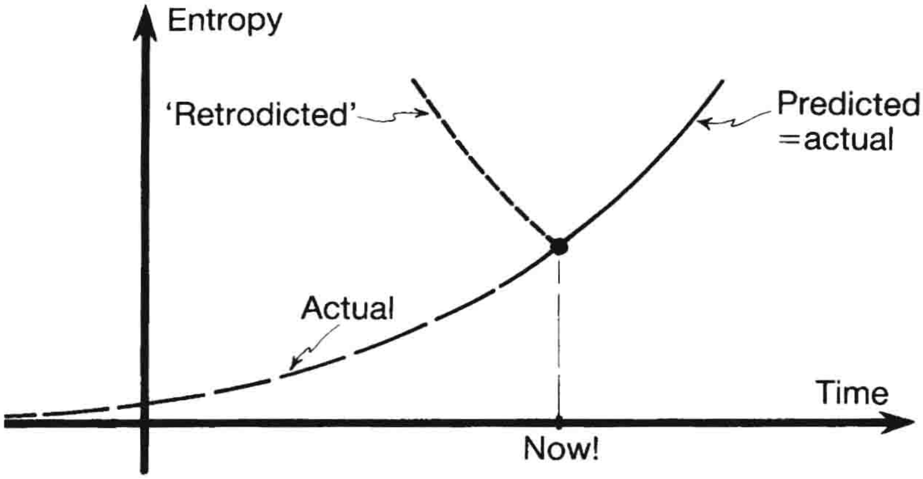


FIGURE 2. The phase-space argument, applied in the reverse direction in time, gives results at variance with observation and with the second law.

smaller regions again once it has reached a larger one. In this way, the entropy continues to increase until ultimately the phase-space point loses itself in the hugest volume of all, namely, that corresponding to thermal equilibrium (see FIGURE 1).

However, this kind of argument can provide only a partial explanation of the second law. In a rough sense, it tells us that given that the universe finds itself in a state of low entropy, then we naturally expect that, in the absence of any other constraints, the entropy of the universe should increase in the future. This argument, though, applies equally well in past directions in time: that is, given that the universe is in a low entropy state at any one time, then it should be natural to believe that the entropy of the universe should increase also in past directions in time, which is in blatant contradiction with the second law (FIGURE 2). Nowhere in our argument have we introduced any distinction between past and future, so it should be clear that this type of argument is not, in itself, capable of deriving the time-asymmetric second law. The essential puzzle that the second law presents us with is that the entropy of the universe gets tinier and tinier the further back into the past we choose to look. Somehow, there must have been an enormous constraint placed on the state of the universe at the beginning of time: at the big bang.

GRAVITATIONAL ENTROPY

What was the nature of this constraint? The first thing to bear in mind is that the degrees of freedom in the matter must be taken to have been completely thermalized in the early stages of the universe. (We need this, in the hot big bang model, in order to get accurate predictions for helium abundance, etc., predictions that indeed closely accord with observation.) Thus, there is nothing about the matter itself that was of "low entropy". The lowness in the entropy lay entirely in the structure of the space-time geometry.

There is a further point that should be emphasized: it was not the fact that the universe was "small" in its early stages that ensured that its entropy was low. A closed universe model that recollapses to a state (in the "big crunch") in which it is as tiny as the one from which it started will still (according to the viewpoint of ordinary physics, where the second law of thermodynamics is maintained right up until the end) achieve an enormous entropy in its tiny final state. The main entropy contained in space-time geometry—gravitational entropy—is not a matter of just total universe-volume.

Let us remind ourselves how gravitational fields can be understood in terms of space-time geometry. In FIGURE 3, I have indicated how the presence of the earth's gravitational field would make itself felt to a system of particles in free-fall. The tidal distortion gives a direct measure of space-time curvature in terms of the equation of geodesic deviation (FIGURE 4; cf. Pirani[12]):

$$D^2q^d = R_{abc}{}^{d}t^a q^b t^c.$$

(Here, t^a is a unit timelike vector field pointing along the particles' world-lines; $D = t^a\nabla_a$ is a propagation covariant derivative along t^a, that is, time-derivative; and q^a is a Lie-propagated "connecting vector", orthogonal to t^a, that labels the displacement of a neighboring particle.) The Ricci tensor, $R_{ab} = R_{acb}{}^{c}$, measures the volume change in the sense that a small collection of particles, held initially at rest, will suffer a displacement (at second order in time, that is, an "acceleration") for which the proportional volume change is measured by $R_{ab}t^a t^b$:

$$D^2\Delta = R_{ab}t^a t^b\Delta.$$

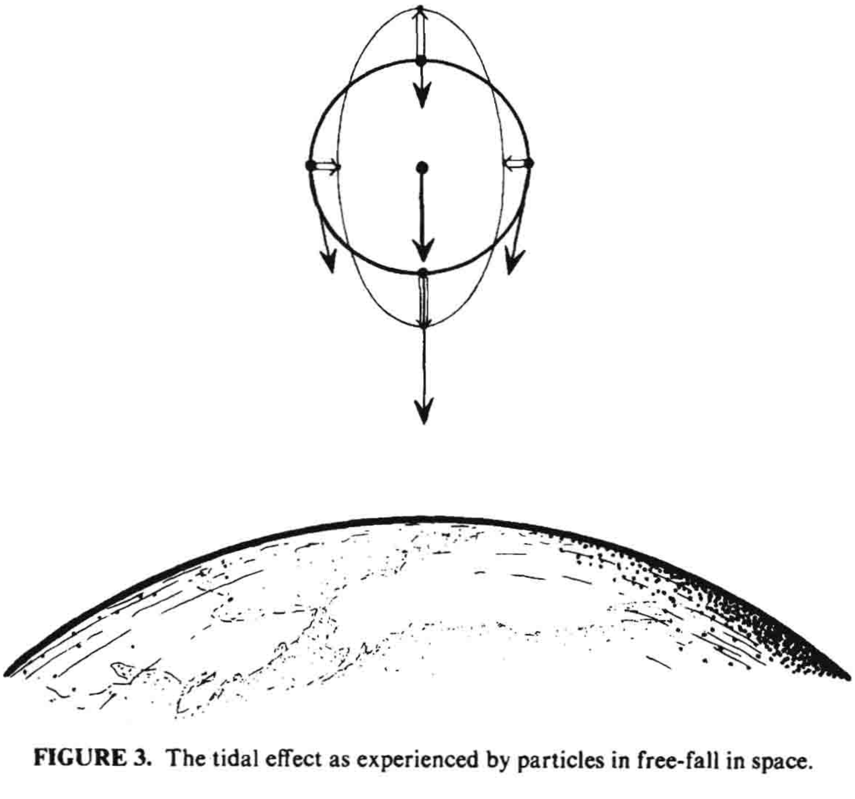


FIGURE 3. The tidal effect as experienced by particles in free-fall in space.

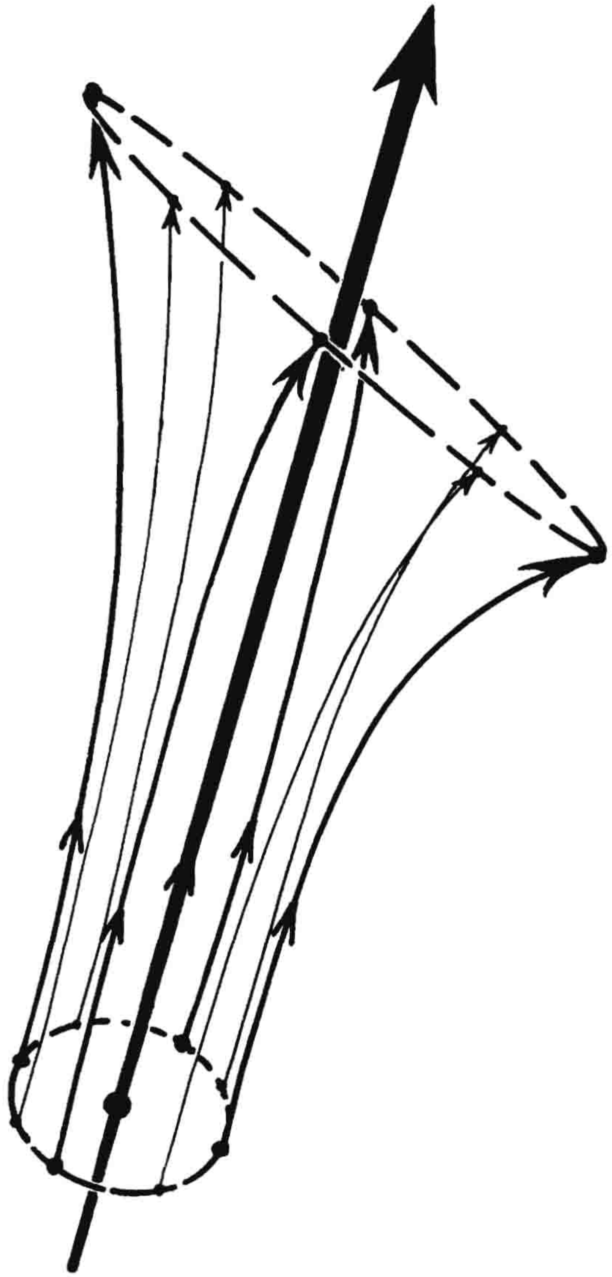


FIGURE 4. Geodesic deviation measures space-time curvature: space-time picture of the tidal effect. (Time is depicted upwards.)

(Here, Δ denotes a small three-volume determined by three independent connecting vectors q^a.) In the situation depicted in FIGURE 3, we are imagining that a sphere of particles is released at rest in a region of vacuum ($R_{ab} = 0$), in which case the initial displacement of the particles relative to a freely falling point at their center is a pure tidal distortion without initial change in volume. This tidal distortion measures the part of the curvature that remains when the Ricci tensor vanishes, namely, the Weyl conformal tensor, C_{abcd}, in the decomposition,

$$R_{abcd} = C_{abcd} + Q_{abcd},$$

of the Riemann tensor, where Q_{abcd} is constructed entirely from the Ricci tensor:

$$Q_{ab}{}^{cd} = 2\,R_{[a}{}^{[c}\,g_{b]}{}^{d]} - \frac{1}{3}\,R g_{[a}{}^{c}\,g_{b]}{}^{d}, \qquad R = R_a{}^a.$$

(Square brackets denote antisymmetrization.) The Weyl tensor describes the part of the curvature that is due to pure gravitational field, whereas the Ricci tensor describes the part that, according to Einstein's equation,

$$R_{ab} = -8\pi G\left(T_{ab} - \frac{1}{2}\,T_c{}^c\,g_{ab}\right),$$

is directly due to the energy-momentum density of matter.

If our sphere of particles were to surround matter rather than just empty space, then there would be a net inward acceleration proportional (approximately) to the total mass surrounded (see FIGURE 5). In rough terms, it is the Ricci tensor (i.e., matter distribution) that determines the volume change and it is the Weyl tensor (i.e., pure gravitational field) that governs the tidal distortion. (This dichotomy is considerably cleaner in the case of light rays rather than for freely falling massive particles; cf. references 13–15. For massive particles, there can be a contribution to the tidal

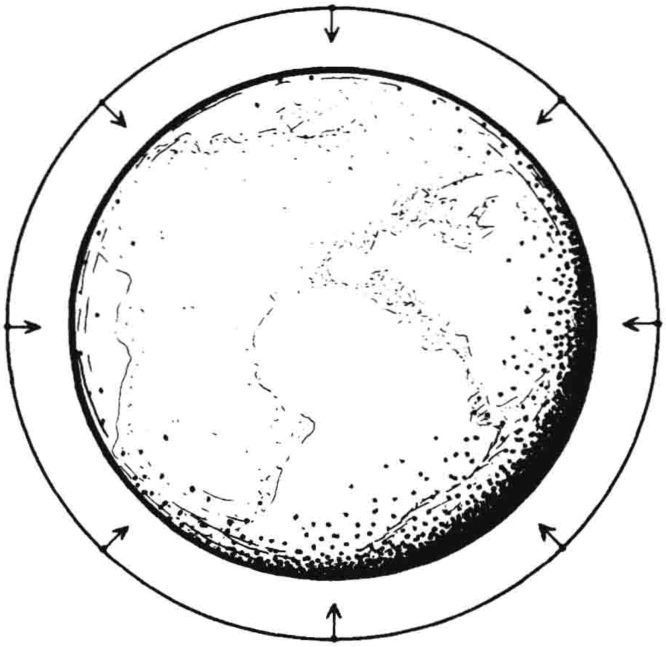


FIGURE 5. When the sphere of particles surrounds matter, there is a net (volume-reducing) acceleration inwards.

CHAPTER 192

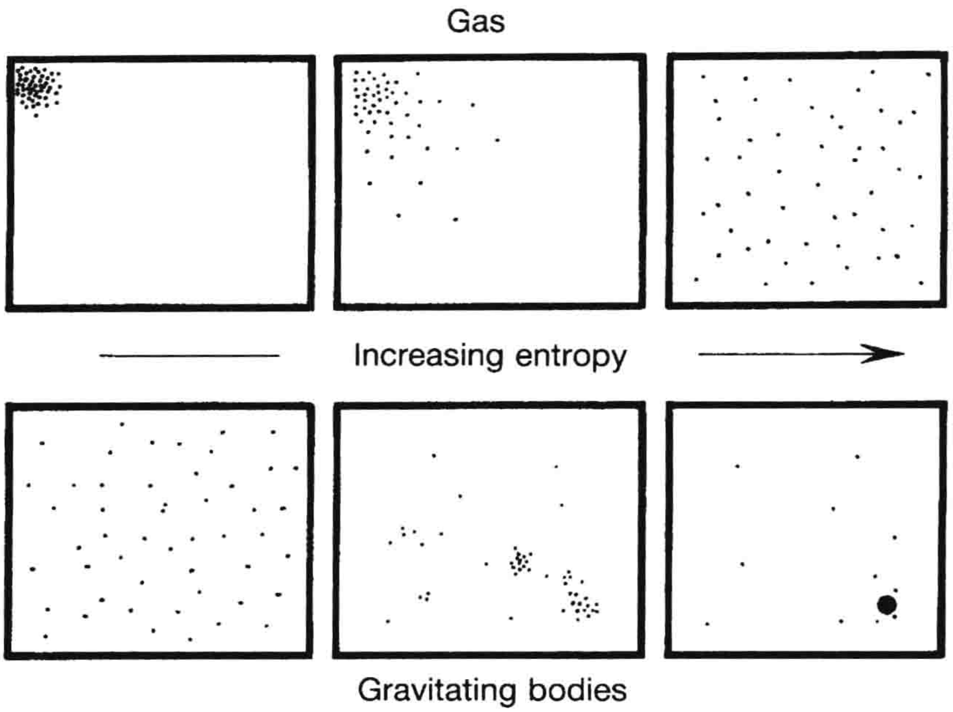


FIGURE 6. With a gas in a box, the maximum entropy state (thermal equilibrium) has the gas distributed uniformly; however, with a system of gravitating bodies, entropy can be increased from the uniform state by gravitational clumping to a black hole finally.

distortion from the presence of Ricci curvature, in addition to that due to Weyl curvature.)

How do we expect the gravitational field to contribute to the entropy of a system? Consider FIGURE 6. In the sequence of three panels at the top, the particles of an ordinary gas are depicted, where increasing entropy is associated with an increasing uniformity in the distribution. In the sequence of three panels at the bottom, we have a system of gravitating bodies and the increase of entropy is associated with an increase of the clumpiness of the distribution, where the maximum entropy state is achieved by most of the material collapsing into a black hole. The low-entropy uniform distribution at the bottom left has a space-time geometry whose curvature is essentially given entirely by the Ricci tensor; in contrast, as the matter clumps and gravitational entropy increases, more and more Weyl tensor appears and, in the extreme case of a black hole, the Weyl curvature diverges to infinity at the singularity.

This seems to suggest that, in some rough way, the higher states of gravitational entropy are to be associated with the presence of greater amounts of Weyl curvature. Unfortunately, there is no mathematical expression known that gives a good measure of the contribution from space-time structure to the entropy of a general system. The famous Bekenstein-Hawking formula for the entropy, S_{bh}, of a black hole of surface area A, namely,[16]

$$S_{\mathrm{bh}} = kc^3A/4G\hbar,$$

is the only clear-cut example of a precise relationship between space-time geometry and entropy. Note the appearance of $\hbar$ in this formula, illustrating the fact that the contribution of space-time geometry to entropy must ultimately be a quantum-mechanical effect. This formula will be of importance for us shortly, but for the moment let us not worry about the absence of a clear-cut expression relating space-time geometry to entropy. Such an expression will not be necessary for what I have to say here. All that we need to note is that (as is illustrated in the lower three panels in FIGURE 6) entropy can be gained from the gravitational field by the clumping of material from an initially uniform distribution with an essentially Weyl-curvature-free initial space-time geometry.

It is important to realize that gravitational clumping is the key ingredient to the entropy imbalance (that is, to the second law of thermodynamics) in the universe as we know it. In FIGURE 7, I have illustrated how it is that we ourselves manage to keep the entropy within us at a necessary low level, by eating low-entropy food obtained (perhaps via animals) from the ability of plants to make use of low-entropy visible-light photons from the sun, the energy of which is ultimately converted into a larger number of entropy infrared photons that escape back into space. For this to be possible, we depend upon the fact that the sun is a hot spot in an otherwise much darker sky. This entropy imbalance in the sky arises essentially because the sun was able to concentrate itself into a single body by gravitational clumping. We are all, directly or indirectly, living off the fact that the hot-spot sun has been created out of a previously uniform

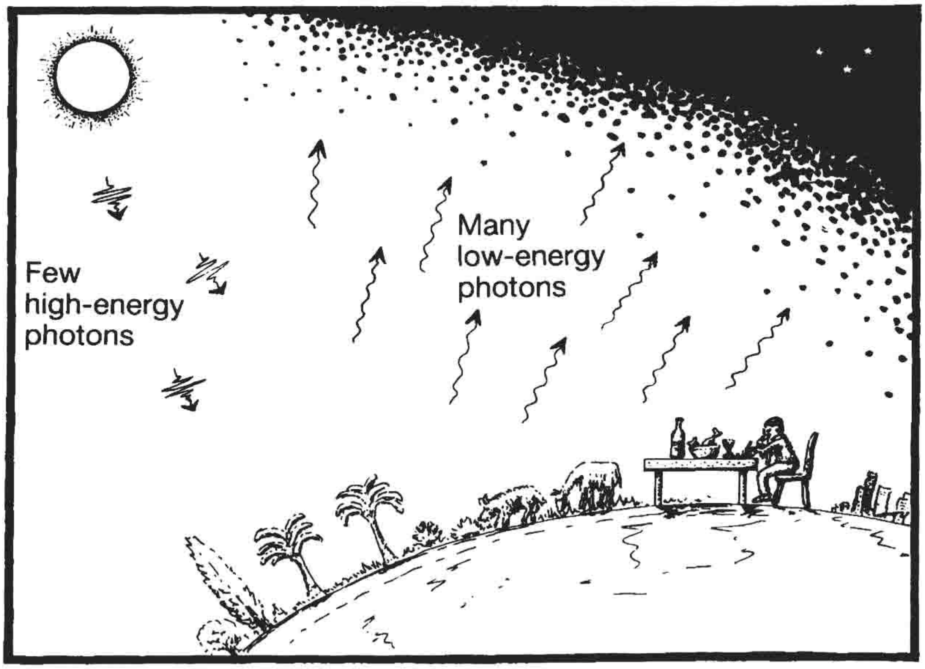


FIGURE 7. The low entropy upon which life on earth depends owes its existence to the fact that the sun is a hot spot in a dark sky, and the sun owes its existence to gravitational clumping.

and even lower-entropy distribution of material. The states of still remarkable low entropy that we find all about us and that give the second law of thermodynamics in the form that we know it come ultimately from the fact that there is a much huger source of low entropy out in space. This low entropy has its origin in the "nonthermalized" gravitational field: the uniform initial state that allows entropy to rise enormously via the mechanism of forming stars through gravitational clumping.

THE LOW ENTROPY IN THE BIG BANG

This uniformity was already present in the big bang. The puzzle of the second law, as explained above, lies in the fact that there must have been an enormous constraint placed on the universe when it was created so that its entropy could start from a very low value and it has been allowed to increase ever since (owing to the absence, presumably, of a similar constraint in the future). This initial constraint was some kind of uniformity condition on the big bang itself. In order to fix our ideas, it will be helpful to consider the situation of a closed universe; let us, for definiteness, take the number of baryons to be 10^{80}. (For a larger number of baryons, the figures obtained below would be even more impressive.) Let us simply take the standard model for such a universe, with its closely Friedmann-Robertson-Walker (FRW) early state, without for the moment asking how this initial state might have arisen (e.g., by inflation?). In FIGURE 8, I have sketched the space-time for such a model. Black holes would be expected to form at some stage (certainly when the universe is still expanding). These would congeal with one another at various times and, ultimately, when the entire universe recollapses in a big crunch, we would expect all the black-hole singularities (with the possible exception of a few that might have disappeared by Hawking evaporation) to unite into one horrendously complicated final singularity.

This final singularity incorporates the enormously high ultimate value of the entropy that our universe model achieves. How high is this value? To put things in perspective, consider the entropy in the present-day background radiation, which is around 10^8 per baryon, in natural units, that is,

$$10^{88}$$

for the entire universe model. (It is best to consider the total entropy rather than just the entropy-per-baryon figure because this allows us to consider that the baryon number need not actually be conserved.) This kind of figure used to impress people by its hugeness because it easily swamps the entropy values for all "ordinary" processes. It is likely, though, that the entropy per baryon of our own universe is now very much larger than this because of the enormous contributions, according to the Bekenstein-Hawking formula, that would arise from black holes. For example, if our universe model were populated by $10^{11}\ M_\odot$ galaxies, each with a $10^6\ M_\odot$ black hole at its core, then it would have an entropy per baryon of 10^{21} and, therefore, a total entropy of

$$10^{101}.$$

We see that this is enormously larger than the value obtained just from the blackbody radiation. We also can use the Bekenstein-Hawking formula to estimate the entropy in the final all-embracing congealing of black-hole singularities that constitutes the big

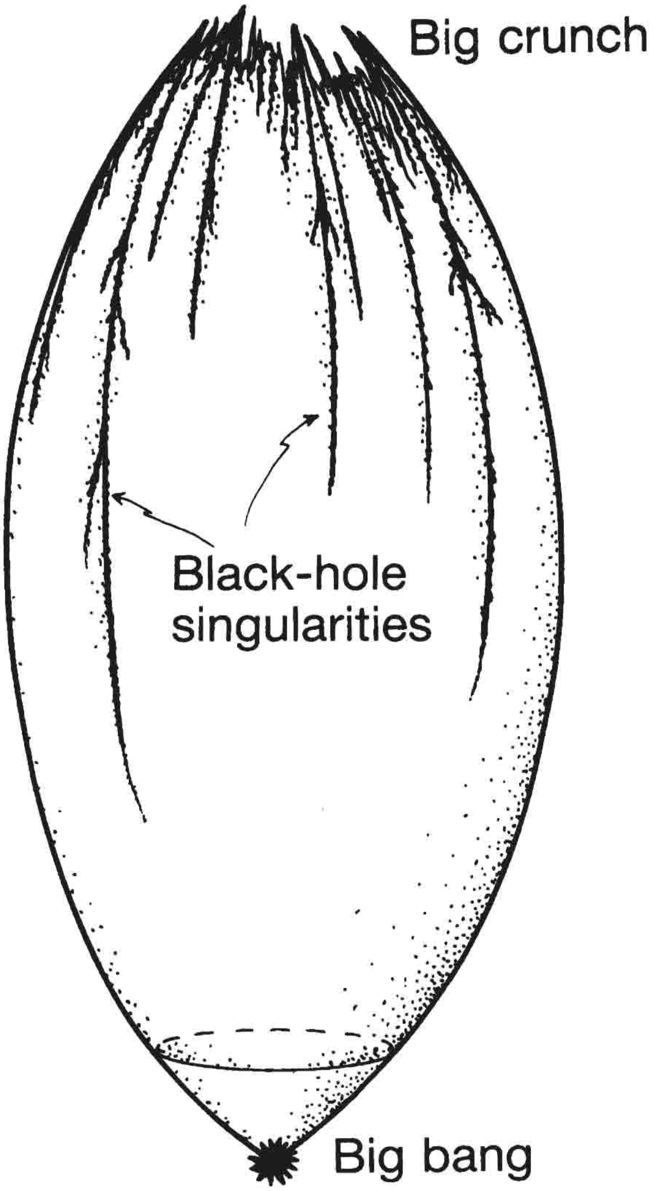


FIGURE 8. A model of a closed universe in which entropy increases by gravitational clumping to black holes and a very complicated big crunch finally.

crunch of our model. We do not go far wrong (certainly as a lower limit) if we use the value for the entropy for a black hole of the mass of all the baryons in the universe for this final entropy:

$$10^{123}.$$

It is entropy values of this kind of size that are available to us in space-time singularity structure. In the face of this fact, we see that the entropy of the background radiation is actually ridiculously tiny. It is not an explanation for a "high" value in initial entropy that we must seek. We must understand why the entropy in the big bang was actually so very small compared with what it might have been. Why was it that the universe (assuming a closed model for the moment) resembles FIGURE 8 rather than the very much more "probable" model depicted in FIGURE 9? For the model of FIGURE 9, I have assumed that the initial "big bang" singularity is now of a high entropy type,

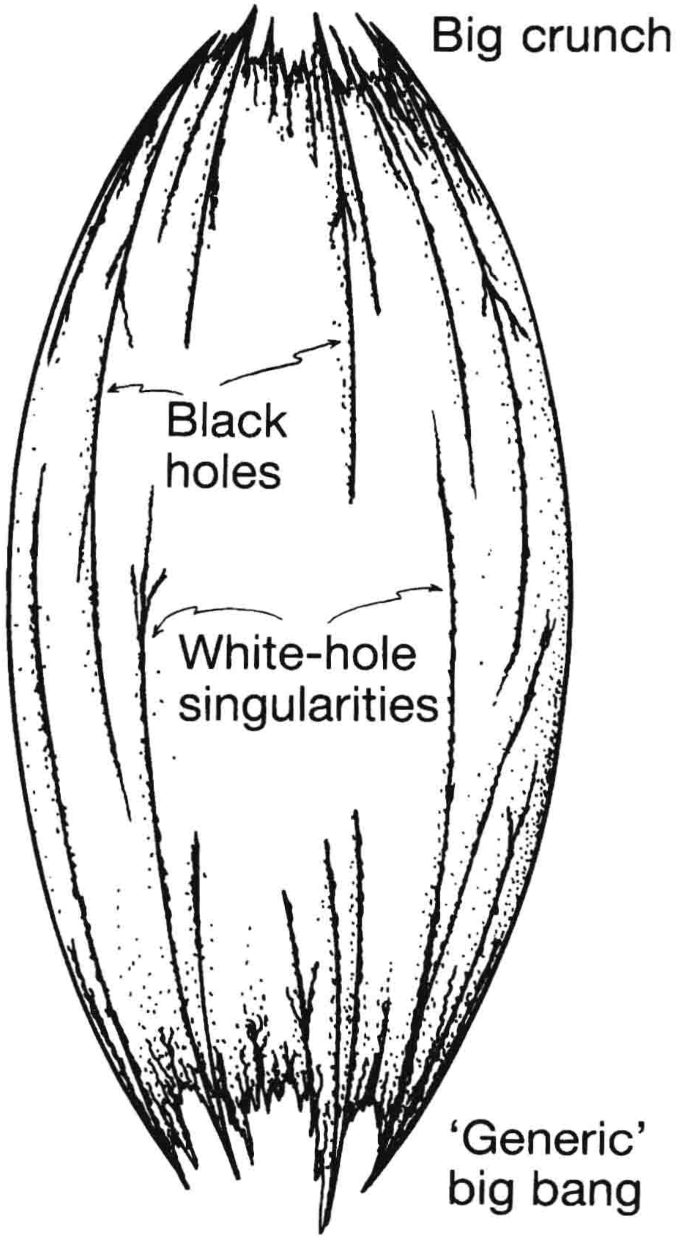


FIGURE 9. A "more probable" closed universe in which the big bang is unconstrained and is as of high an entropy as the big crunch.

like the big crunch of the more realistic model of FIGURE 8. The model of FIGURE 9 would have no second law of thermodynamics because its entropy starts off with the same enormously high entropy value that it ends up with. Thus, our question implicitly contains a second question: Why is there a second law of thermodynamics at all?

THE WEYL CURVATURE HYPOTHESIS

Let us phrase our question in terms of phase-space volumes. The puzzle that needs to be explained is as follows: Why was the initial state of the universe—the big bang—constrained to such a very tiny fraction of the whole volume of phase-space, namely,

$$\text{one part in } 10^{10^{123}},$$

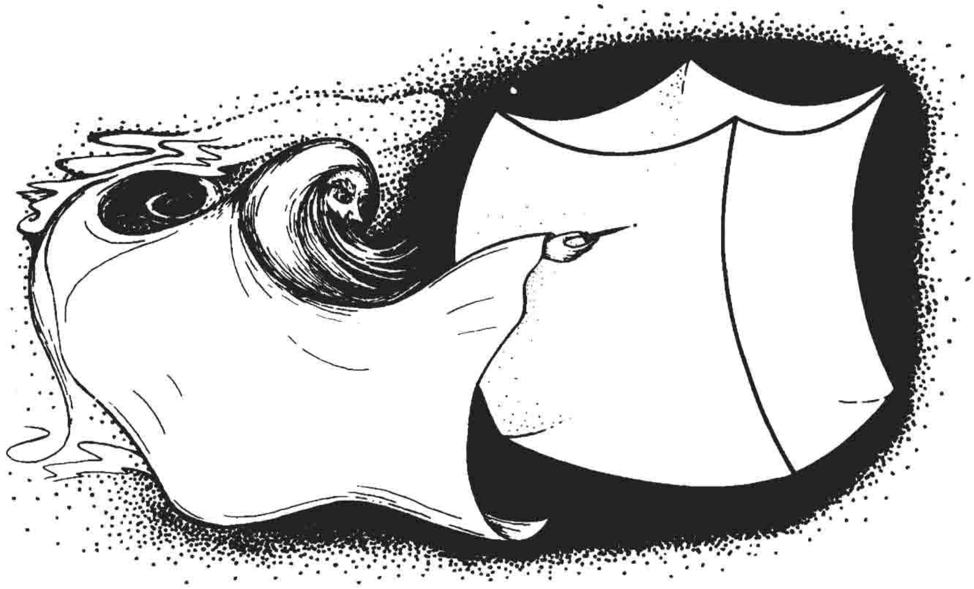


FIGURE 10. The Creator locating the tiny region of phase-space—one part in $10^{10^{123}}$—needed to produce a 10^{80}-baryon closed universe with a second law of thermodynamics in the form we know it.

in the case of our particular universe model? (Recall the logarithmic relation between entropy and phase-space volume.) This question is made graphic in FIGURE 10, where we imagine the Creator, armed with a pin, aiming for this absurdly tiny region of phase-space that is necessary in order to provide us with a universe of the uniformity needed for a second law of thermodynamics in the form that is familiar to us.

Several points should be made here. In the first place, our assumption of a closed universe (with, say, 10^{80} baryons) is not essential to the argument, though it enables us to consider that the phase-space available to the Creator might have finite volume. For an infinite universe volume, the phase-space would have to have infinite volume. The region that the Creator would have to aim for would have infinite volume as well, but it would still be only a tiny fraction of the entire volume—now a vanishingly small fraction, in fact. The larger the universe model, the greater would be the precision required on the part of the Creator.

In the second place, the suggestion, as is sometimes put forward, that it is the anthropic principle that fixes the conditions of our universe—and so we have a suitable second law of thermodynamics—simply will not do. The phase-space precision that would be required in order to create our entire solar system out of random particle collisions, with an earth teeming with life just as we know it, is nothing by comparison with a figure like $1/10^{10^{123}}$. (A precision to $1/10^{10^{50}}$ would be far more than enough, for example.) The anthropic principle can provide no explanation for $10^{10^{123}}$.

Third, people sometimes seem to worry (cf. Turner,[17] for example) that examining the degree of specialness of the universe "at the present time" and "at the beginning" might be two quite different things. Consequently, the above discussion might, in some essential way, be inappropriate. However, it should be pointed out that Liouville's theorem tells us that phase-space volumes do not change in the course of dynamical evolution, so it makes no difference with regard to the questions of "fractions of phase-space volumes" whether we base our discussion on the universe "now" or "at the beginning". If the Creator's choice was made purely randomly, it would not matter at what "time" the choice was actually made. (Indeed, we can regard the phase-space as representing entire universe-histories if we wish, rather than states of the universe at particular times, and the phase-space volume arguments will apply equally well.) However, the coarse-graining does change with time. If the Creator's choice is governed by considerations other than pure randomness, then it can make all the difference in the world "when" this choice is made.

It should also be remarked that the precision that is needed in order to provide us with a big bang singularity of the type required may be viewed as a ruling out of white holes in the initial state (see FIGURE 9 again). White holes are antithermodynamic objects. (They have horizon sizes that decrease with time and can generally violate the second law of thermodynamics.) There do not seem to be any white holes present in our universe; however, it should be remarked that it is not really an observational fact that white holes do not exist, but a profound theoretical requirement of consistency with the second law of thermodynamics. The absence of a proliferation of white holes in our universe can be used as an argument against a "time-symmetric" viewpoint that is heard from time to time: that the entropy direction should reverse itself in the collapsing phase of a closed universe. Because there will be many black holes present in the final collapsing phase, there would, in this time-symmetrical view, have to have been as many white holes present in the expanding phase. Their absence adds to other impressive arguments that we can make against this time-symmetric view (cf. Davies[18]).

What is it then that forces our Creator's pin into such a tiny region of phase-space? If we believe that this is a question for science rather than unknowable mystery or a Creator's whim, then we must seek an explanation within the laws of physics. Because space-time singularity structure is at issue here, it is to quantum gravity that we must presumably turn for this explanation. Unfortunately, there is not yet any comprehensive or believable theory of quantum gravity; however, when one comes to light, among its implications would seem to be a constraint on the big bang to be a singularity of very uniform type. To imply the second law, this constraint would need to apply time-asymmetrically, and I envisage that any space-time singularities of initial type (i.e., big bang or white hole) must be constrained so that the Weyl curvature vanishes in the immediate vicinity of the singularity, whereas singularities of final type (i.e., big crunch or black hole) would not be constrained in this way. I call this characterization of space-time singularities the Weyl curvature hypothesis. For the moment, it is, indeed, merely a hypothesis,[b] but it has important implications. Essentially, these are the uniformity (homogeneity, isotropy) of the initial state of the universe (where I assume thermalization of the matter content of the early universe), the absence of white holes (and of mini–black holes), and the second law of thermodynamics in the form that we know it. It is the Weyl curvature hypothesis (presumably an implication of the true quantum gravity) that forces the Creator's pin into the required tiny region of phase-space.

CAN AN EXPLANATION BE OBTAINED FROM INFLATION?

Proponents of the inflationary view would presumably seek to explain this initial uniformity as an effect of an enormous expansion of a previously tiny region of space-time. There are, however, profound problems inherent in an explanation solely in such terms. If it is supposed that the initial space-time singularity that quantum gravity presents us with is, in some sense, "generic", then this would mean a randomly chosen initial point in the phase-space: giving an initial singularity that is the time-reverse of a generic big-crunch singularity (as I have tried to depict in FIGURE 9). This is very far from the type of situation that has been studied in the analysis of the inflationary mechanism. Virtually all detailed investigations so far have taken the FRW models as their starting point, which, as we have seen, totally begs the question of the enormous number of degrees of freedom available in the gravitational field (initial freedom in the choice of Weyl tensor).

Suppose we believe that inflation could actually take place, starting from such a generic initial singularity (which seems to me to be in any case extremely doubtful). What would the result of such an inflation be? Would we expect, as all proponents of the inflationary viewpoint seem to assume, that a spatially very uniform universe would be the result? This seems to me to be unlikely. There is some evidence that chaotic behavior might be a feature of generic space-time singularities (cf. references 20–22, although those discussions were dependent upon assumptions of spatial homogeneity). If so, and hence fractal-like behavior is to be expected on the small scale, then an inflationary expansion will be of no use at all in producing a spatially uniform universe. An apparently even more serious problem arises when we take the "quantum foam" implications of (some versions of) quantum gravity into account.[23] According to this picture, the topological structure of space-time gets very complicated at the scale of the Planck length (10^{-33} cm), where there would be a proliferation of "wormholes" and other types of topological peculiarity. With an inflation factor of, say, 10^{50}, would we not expect to see the effects of such topological peculiarities expanded out to the scale of light-years?

---

[b] The proposal of Hartle and Hawking[19] might possibly give theoretical substance to a quantum gravity origin for space-time singularities having vanishing Weyl curvature. However, as it stands, their proposal lacks the time-asymmetric ingredient that is essential, in my opinion, if the second law of thermodynamics is to find explanation.

It seems to me that the idea of trying to obtain a universe in the form that we know it by applying time-symmetric physics to a generic unconstrained initial state is basically misconceived. The situation with regard to inflation seems to me to be basically the same as it was with the older idea of "chaotic cosmology".[20,21] The second law of thermodynamics arises because there was an enormous constraint (of a very particular kind) placed on the universe at the beginning of time, giving us the very low entropy that we need in order to start things off. Choosing a generic initial state amounts to starting the universe off at a maximum entropy state, which is the wrong way around.

In addition, it should be made clear that the physical theory that has (so far) been used as the input for the inflationary picture is indeed entirely time-symmetric. The mechanism of symmetry-breaking is time-asymmetric, of course, but the time-asymmetric input here comes entirely from the second law of thermodynamics. This really begs the question because it is precisely the second law itself that needs explaining and presents the greatest puzzle as to the extraordinary uniformity of the initial state.

Finally, I should address the suggestion sometimes made by inflationists that perhaps some very tiny part of an initially chaotic state takes part in an inflationary expansion only when its structure is smooth enough, with this expansion resulting in our entire observable universe. Believers in such a picture should consider their position in the light of the difficulties raised above, particularly in relation to my discussion of the anthropic principle. Beyond this, in the words of Börner,[6] "It seems difficult to introduce a quantitative aspect into such a romantic world view."

REFERENCES

1. GLINER, E. B. 1965. Sov. Phys. JETP **22:** 378.
2. SATO, K. 1981. Mon. Not. R. Astron. Soc. **195:** 198, 487; **99b:** 66.
3. GUTH, A. H. 1981. Phys. Rev. **D23:** 347.
4. LINDE, A. D. 1982. Phys. Lett. **108B:** 389.
5. ALBRECHT, A. & P. J. STEINHARDT. 1982. Phys. Rev. Lett. **48:** 1220.
6. BÖRNER, G. 1988. The Early Universe—Facts and Fiction. Springer-Verlag. Berlin/New York.
7. PENROSE, R. 1986. Book review of *The Very Early Universe* [G. W. Gibbons, S. W. Hawking & S. T. C. Siklos, Eds. Cambridge University Press (paperback edition). London/New York]. Observatory Magazine.
8. PENROSE, R. 1988. Fundamental asymmetry in physical law. Proc. of Symposia in Pure Mathematics **48**.
9. PENROSE, R. 1979. Singularities and time-asymmetry. *In* General Relativity: An Einstein Centenary System. S. W. Hawking & W. Israel, Eds. Cambridge University Press. London/New York.
10. PENROSE, R. 1981. Time-asymmetry and quantum gravity. *In* Quantum Gravity 2. D. W. Sciama, R. Penrose & C. J. Isham, Eds. Oxford University Press. London/New York.
11. PENROSE, R. 1989. The Emperor's New Mind. Oxford University Press. London/New York.
12. PIRANI, F. A. E. 1957. Phys. Rev. **105:** 1089.
13. PIRANI, F. A. E. & A. SCHILD. 1961. Bull. Acad. Pol. Sci. Ser. Sci. Math. Astron. Phys. **9:** 543.

14. SACHS, R. K. 1962. Proc. R. Soc. London **A270:** 103.
15. PENROSE, R. 1966. General-relativistic energy flux and elementary optics. *In* Perspectives in Geometry and Relativity. B. Hoffmann, Ed. Indiana University Press. Bloomington, Indiana.
16. HAWKING, S. W. 1975. Commun. Math. Phys. **43:** 199.
17. TURNER, M. S. 1987. Inflation in the universe, circa 1986. *In* General Relativity and Gravitation—Proceedings of the 11th International Conference on GRG. M. A. H. MacCallum, Ed.: 238–239. Cambridge University Press. London/New York.
18. DAVIES, P. C. W. 1974. The Physics of Time Asymmetry. Surrey University Press. Surrey, England.
19. HARTLE, J. B. & S. W. HAWKING. 1983. Phys. Rev. **D31:** 1777.
20. MISNER, C. W. 1968. Astrophys. J. **151:** 431.
21. MISNER, C. W. 1969. Phys. Rev. Lett. **22:** 1071.
22. BELINSKII, V. A., I. M. KHALATNIKOV & E. M. LIFSHITZ. 1970. Adv. Phys. **19:** 523.
23. WHEELER, J. A. 1962. Geometrodynamics. Academic Press. New York/London.

193
LIGHT RAYS NEAR i^0: A NEW MASS-POSITIVITY THEOREM

Twistor Newsletter **30** (1990). Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. Chapman and Hall/CRC, London, 1995

This is a genuine argument for mass positivity in general relativity, but the geometrical reasoning presented pictorially here would be hard to formalize. A closely related article addressing this issue, written in collaboration with Raphael Sorkin and Eric Woolgar, has still not been put into a satisfactory finalized form. The relevant geometrical conditions at spacelike infinity i^0 are studied in Chapter 199.

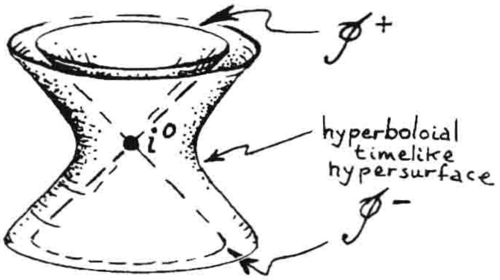
20 Light rays near i^0: A new mass-positivity theorem. *Twistor Newsletter* **30** (1990). Also published in *Further Advances in Twistor Theory*, Vol. 2. Eds L. J. Mason and L. Hughston. 1995

§II.3.10 **Light Rays near i^0: A new Mass-Positivity Theorem** *by R.Penrose* ($\mathbb{TN}$ 30, June 1990)

As was emphasized by Helmut Friedrich in a recent survey talk [given at the E.T. Newman birthday conference, Pittsburgh, May 1990], one of the major problems obstructing a proper understanding of asymptotically flat space-times is the lack of a completely satisfactory geometrical framework for analysing i^0 (the conformally singular point at spacelike infinity). Perhaps twistor theory can make a significant contribution here. Whereas partial understandings have come through the work of Ashtekar-Hanson, Geroch, Beig, Sachs, Sommers and many others, there remains a distinct conceptual awkwardness, and as yet there is no elegant geometrical description to match—and to unite with—that of $\mathscr{I}^{\pm}$. The 'spi' construction of Ashtekar-Hansen (1978) seems to come closest and there are indeed some relations to twistor theory. In Minkowski space $\mathbb{M}$, the points of spi can be identified with timelike hyperplanes (i.e. timelike 3-quadrics through i) and these are represented by (proportionality classes of) skew twistor $S^{\alpha\beta}$ subject to $S^{\alpha\beta}I_{\alpha\beta} = 0$ and the "reality condition" $\bar{S}_{\alpha\beta}I^{\alpha\beta} = I_{\alpha\beta}S^{\gamma\beta}$. Such objects find their place in the *moment sequence* (see Penrose & Rindler 1986, pp 94, 96 and §II.4.2-3 in this volume). The points of $\mathscr{I}$ arise when $S^{\alpha\beta}$ is *simple* (and so has the form $Z^{[\alpha}I^{\beta]\gamma}W_{\gamma}$).

Light rays near i^0 A new mass-positivity theorem. *Twistor Newsletter* **30** (1990). Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. 1995

I shall postpone, until a later date, a detailed discussion of the geometry involved here. In any case, we cannot expect to capture the structure of i^0 that incorporates the total mass-momentum and angular momentum of the system using merely Minkowski-space twistors. Possible schemes for involving more general types of twistor would be to use 2-surface twistors in the manner of Shaw, or perhaps to use hypersurface twistors, e.g. for a hyperboloidal timelike hypersurface that approximates $i^0 \cup \mathscr{I}^+ \cup \mathscr{I}^-$.



Of more immediate physical relevance would be to study the behaviour of light rays near i^0. There is a clear connection with twistor theory here, but for the moment I shall just show how this study can be used to give a new and perhaps comparatively simple proof of total energy positivity in general relativity (cf. Schoen & Yau (1979) and Witten(1981)):

THEOREM. *If $\mathcal{M}$ is asymptotically simple and satisfies the strong null convergence condition, then it cannot have a well-defined negative mass at i^0.*

N.B. By the *'strong' null convergence condition* (SNCC) I shall mean that every endless null geodesic contains a pair of conjugate points. This is a consequence of the three properties:

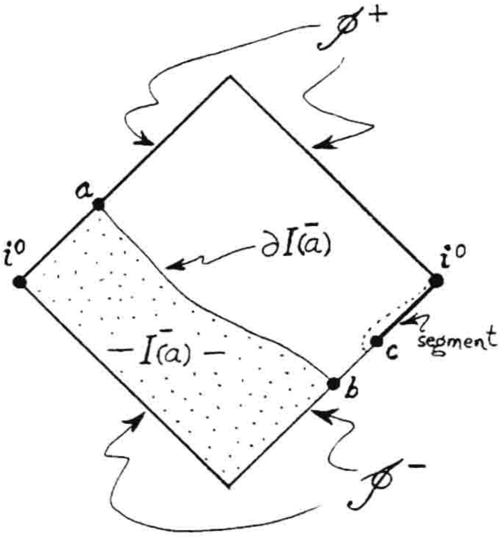
(1) $R_{ab}n^a n^b \leqslant 0$ for all null vectors n^a (i.e. with Einstein's equations, and my sign conventions, $T_{ab}n^a n^b \geqslant 0$),

(2) null geodesic completeness, and

(3) the genericity condition: $k_{[c}R_{a]bc[d}k_{f]}n^b n^c \neq 0$ somewhere along each null geodesic (n^a the tangent). [see Hawking & Penrose, Proc. Roy. Soc. (Lond.) **A314** (1970) 529.]

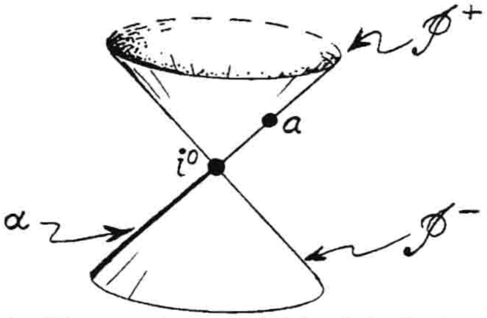
Outline of proof. The idea is to assume that the mass at i^0 *is* negative and that SNCC holds, and then to derive a contradiction. Let $a \in \mathscr{I}^+$ and consider $I^-(a)$. (All sets are in $\bar{\mathcal{M}} = \mathcal{M} \cup \mathscr{I}^+ \cup \mathscr{I}^-$; here $I^-(a)$ is to include its limit points on $\mathscr{I}^-$ (interior limit points only, so that $I^-(a)$ remains open); "$\partial I^-(a)$" is supposed to be composed of the points of $\partial I^-(a)$ in $\mathcal{M}$, together with the limit points thereof in $\bar{\mathcal{M}}$.) Let us see what would happen if it could be shown that some segment of a generator of $\mathscr{I}^-$, from a point $c \in \mathscr{I}^-$ to i^0 has a neighbourhood in $\bar{\mathcal{M}}$ that does not meet $I^-(a)$, as follows if any $c \in \mathscr{I}^-$ has a neighbourhood not meeting $I^-(a)$.

22 Light rays near i^0: A new mass-positivity theorem. *Twistor Newsletter* **30** (1990). Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. 1995

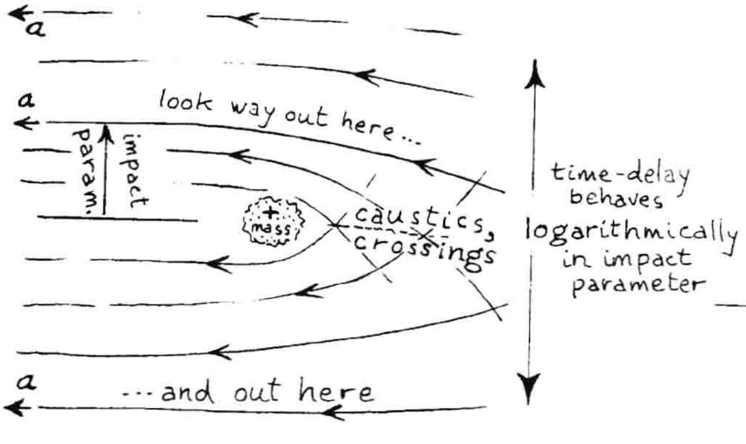


In that case $\partial I^-(a)$ would meet $\mathscr{I}^-$ transversally in some (non-vacuous) set $\mathcal{B}$ $(= \partial I^-(a) \cap \mathscr{I}^-)$. Let $b \in \mathcal{B}$. By standard theorems (since $\mathcal{M}$ is globally hyperbolic, as follows from its asymptotic simplicity) there must be a null geodesic from b to a, lying on $\partial I^-(a)$, with no pair of conjugate points between b and a, which contradicts SNCC.

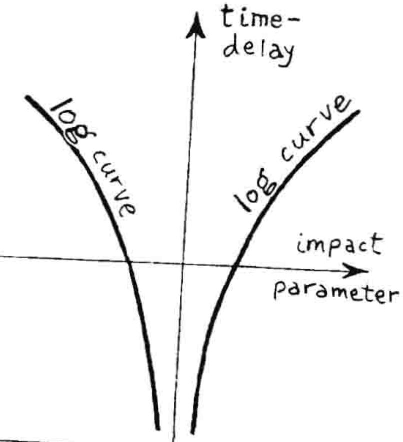
In fact there is *no* such segment (or point c), but what *does* happen, with negative mass at i^0, is almost as effective. Let us examine the generators of the past light cone $\mathcal{A}$ of a (we have $\partial I^-(a) \subset \mathcal{A}$), particularly in the neighbourhood of that generator α of $\mathscr{I}^-$ that is 'diametrically opposite' to a (i.e. α extends through i^0 to become the generator of $\mathscr{I}^+$ containing a).



We shall be mainly concerned with generators of $\mathcal{A}$ "close" to that particular generator of $\mathscr{I}^+$ on which a lies (i.e. α extended), and with their intersection with $\mathscr{I}^-$. To assist us in picturing this situation, consider first the case of *positive* mass at i^0, to see why there is no conflict with SNCC. Not only are the light rays deflected inwards as they pass the source, but there is also a time-delay that behaves logarithmically in the impact parameter (distance of "closest approach"). This has no natural zero, so the larger the impact parameter, the larger the values of advanced time that will eventually be encountered.

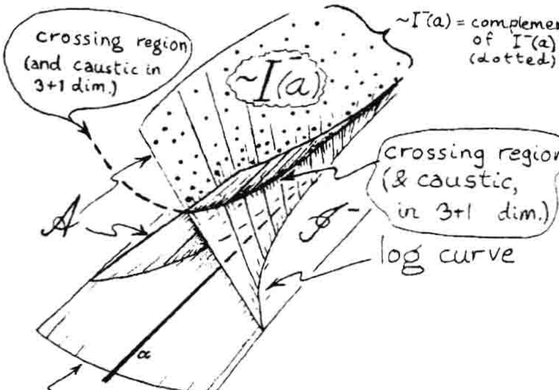

a
a
look way out here...
impact param.
mass
caustics, crossings
a
...and out here
time-delay behaves logarithmically in impact parameter

Spatial picture

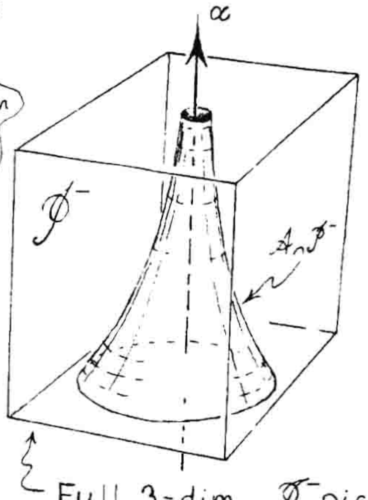


Graph of time-delay

This phenomenon was described in detail in Penrose (1980). We find that the whole of $\mathscr{I}^-$ is contained in $I^-(a)$. (Hence $\partial I^-(a) \cap \mathscr{I}^- = \varnothing$, so there is no contradiction with SNCC.) To see how this comes about, examine $\mathcal{A}$ and $\partial I^-(a)$ near α. We find


crossing region (and caustic in 3+1 dim.)
~I⁻(a) = complement of I⁻(a) (dotted)
~I(a)
𝒜
𝒜
crossing region (& caustic, in 3+1 dim.)
𝒥⁻
log curve
α

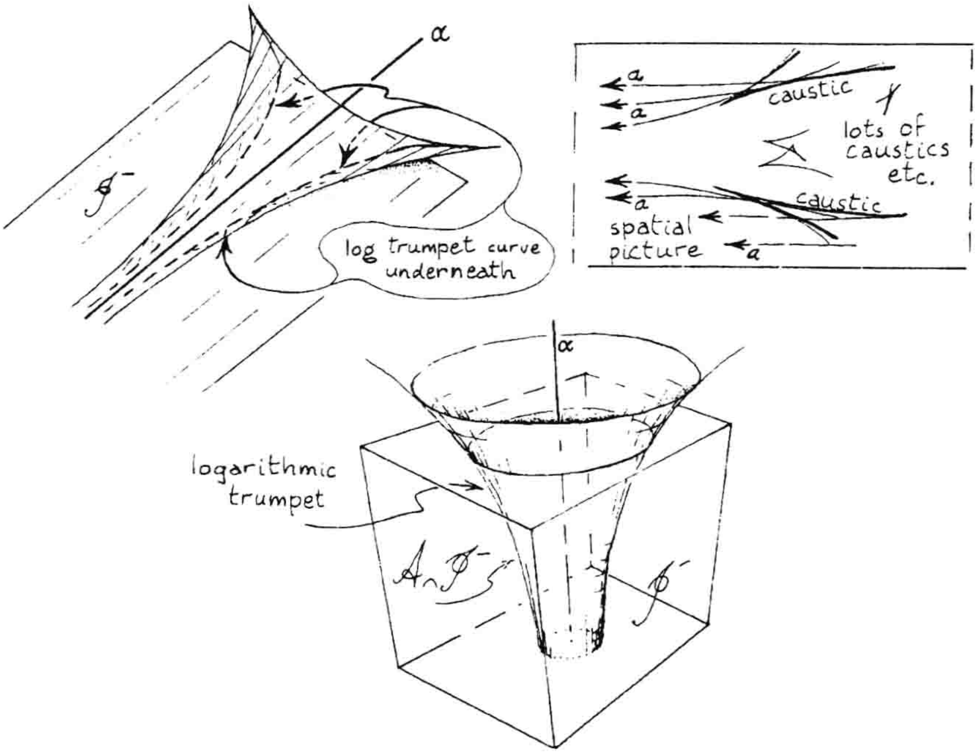
Reduced 2-dim. picture



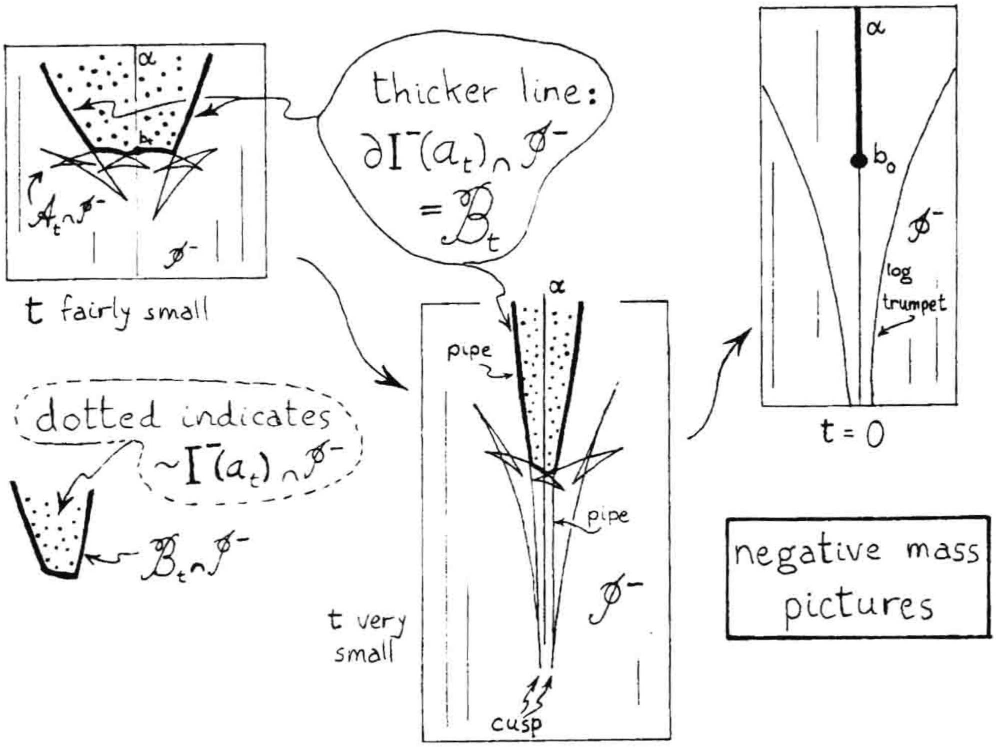
Full 3-dim. $\mathscr{I}^-$ picture
($\mathcal{A} \cap \mathscr{I}^-$ is logarithmic trumpet)

Now consider *negative* mass. The picture is like the one above but the other way up:

24 Light rays near i^0: A new mass-positivity theorem. *Twistor Newsletter* **30** (1990). Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. 1995

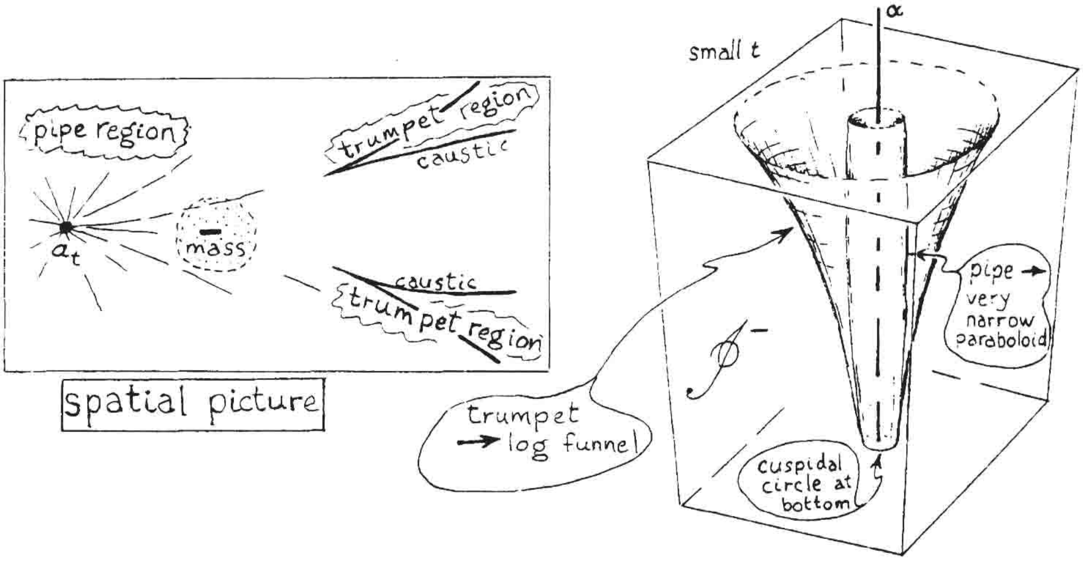

α
$\mathscr{I}^-$
log trumpet curve underneath
a
a
caustic
lots of caustics etc.
a
caustic
spatial picture
a
α
logarithmic trumpet
$\mathscr{A}\cap\mathscr{I}^-$
$\mathscr{I}^-$

What's not so clear from this picture is the fact that $\mathscr{I}^- \subset \overline{I^-(a)}$ $(= I^-(a) \cup \partial I^-(a))$, as in the positive and zero mass cases $\partial I^-(a)$ does not intersect $\mathscr{I}^-$ transversally and we have no immediate contradiction with SNCC. To see what *does* happen, consider the following diagram.


α
b_t
$\mathscr{A}_t\cap\mathscr{I}^-$
$\mathscr{I}^-$
t fairly small
thicker line: $\partial I^-(a_t)\cap\mathscr{I}^- = \mathscr{B}_t$
dotted indicates $\sim I^-(a_t)\cap\mathscr{I}^-$
$\mathscr{B}_t\cap\mathscr{I}^-$
α
pipe
pipe
$\mathscr{I}^-$
t very small
cusp
α
b_0
$\mathscr{I}^-$
log trumpet
t = 0
negative mass pictures

In this diagram we have a smooth family of points $a_t (t \geqslant 0)$ lying on (say) a null ray terminating at a, where $a = a_0$. For each $t > 0$, $\partial I^-(a)$ intersects $\mathscr{I}^-$ transversally in a (not necessarily smooth) cut of $\mathscr{I}^-$. The point $b_t = \alpha \cap \partial I^-(a_t)$ lies on a null ray on $\partial I^-(a_t)$ free of conjugate points (except possibly at end-points). As $t \to 0$ the intersection $\mathcal{A}_t \cap \mathscr{I}^-$ (where $\mathcal{A}_t$ is past light cone of a_t) looks like the above diagram. The region $\mathcal{B}_t = \partial I^-(a_t) \cap \mathscr{I}^-$ moves continuously with t, finally converging on a future-endless segment of α with past end-point $b_0 = b$.

One must show that the part of $\mathcal{A} \cap \mathscr{I}^-$ that lies away from the logarithmic trumpet and the narrower pipe within it actually stabilizes as $t \to 0$ so that b_t attains a limit. The pipe and trumpet come from the resemblance of $\mathcal{M}$'s i^0 to that of a negative mass Schwarzschild solution as can be seen from examining the spatial picture. Having b_0, we find the null ray on $\partial I^-(a_0)$ from b_0 to a_0 which must be free of conjugate points: the desired contradiction with SNCC.



No doubt the argument can be strengthened in various ways (e.g. allowing conjugate points at $\mathscr{I}^{\pm}$ or the mass to be zero). Further details will appear elsewhere.

References

Ashtekar, A.& Hansen (1978) J. Math. Phys. **19**, 1542.

Hawking, S.W. & Penrose, R. (1970) Proc. Roy. Soc. Lond. **A314** 529.

Penrose, R. (1980) in the Taub Festschift volume,(F.Tipler ed.).

Penrose, R. (1986) *Spinors and space-time*, Vol.2, CUP.

Schoen & Yau, (1979) Phys. Rev. Lett. **43** (1979) 1457, and Commun. Math. Phys. **65**, 45.

Witten, (1981) Commun. Math. Phys. **80**, 381.

194

EXTENDED REGGE TRAJECTORIES: 12 YEARS AFTER

Twistor Newsletter **30** (1990). With S. T. Tsou

This looks back at our observation concerning the extendibility of Regge trajectories to negative *j*-values, made in Chapter 95. I am not aware that the issue has ever been picked up seriously by the particle-physics community. See also Chapter 197, which explains, on pp. [10] and [11], why we had an interest in this phenomenon.

28 Extended Regge trajectories: 12 years after. *Twistor Newsletter* **30** (1990). With S. T. Tsou

We wish to give here an extremely brief summary of the status of extended Regge trajectories[1]. We shall present a more detailed report after the new 1990 Particle Data Tables are out.

There are 10 resonances (5 mesons, 5 baryons) whose spin-parity (J^P) were unknown and predicted in our paper in 1978. Of these ten, five (four with confidence) were assigned negative j values. Nine of these J^P assignments are confirmed by the 1988 Tables,[2] and the remaining one is wrong. In the table below we set out the comparison:

1976		Predicted $J^P(j)$	1988	
$D(1285)$	$0^+ (A)^+$	$1^+ (1)$	$f_1(1285)$	$0^+(1^{++})$
$E(1420)$	$0^+ (A)^+$	$1^+ (-2)$	$f_1(1420)$	$0^+(1^{++})$
$A_4(1900)$	$1^-(?)$	$2^-(-3)$	$\pi_2(2100)$	$1^-(2^{-+})$
$\rho(2100)$	$?$	$3^+(-4)$	$\rho_3(2250)$	$1^+(3^{--})$
$K_N(1700)$	$\frac{1}{2}(?)$	$1^-(-2)$	$K^*(1715)$	$\frac{1}{2}\ (1^-)$
$N(2650)$	$\frac{1}{2}(?^-)$	$\frac{13}{2}^+(\frac{13}{2})$	$N(2700)$	$\frac{1}{2}\ (\frac{13}{2}^+)$
$\Delta(2160)$	$\frac{3}{2}(?)$	$\frac{7}{2}^-(\frac{7}{2})$	$\Delta(2200)$	$\frac{3}{2}\ (\frac{7}{2}^-)$
$\Delta(2850)$	$\frac{3}{2}(?^+)$	$\frac{15}{2}^+(-\frac{17}{2})$	$\Delta(2950)$	$\frac{3}{2}\ (\frac{15}{2}^+)$
$\Lambda(2350)$	$0\ (?)$	$\frac{9}{2}^+(\frac{9}{2})$	$\Lambda(2350)$	$0\ (\frac{9}{2}^+)$
$\Xi(1820)$	$\frac{1}{2}(?)$	$\frac{3}{2}^-(\frac{3}{2})$	$\Xi(1820)$	$\frac{1}{2}\ (\frac{3}{2}^-)$

We note that not only have data changed considerably during the 12 years, but all mesons have been renamed!

Roger Penrose
Tsou Sheung Tsun.

References

1. Roger Penrose, George Sparling and Tsou Sheung Tsun, J. Phys A 11 (1978) L231–L235.
2. Particle Properties Data Booklet, April 1988.

195

PRÉCIS OF *THE EMPEROR'S NEW MIND*: CONCERNING COMPUTERS, MINDS AND THE LAWS OF PHYSICS

Behavioral and Brain Sciences **13** (1990), no. 4, 643–55

Producing this summary of my book *The Emperor's New Mind* (Oxford University Press, 1989) was part of the 'open peer' procedure required by this journal. This précis was followed, in the journal, by 37 reviews of the book (the Open Peer Commentary), many of which were highly critical of the position that I was taking. My responses are provided in Chapter 196. The summary given here is fairly comprehensive, and it probably gives a fair impression of what my book was about.

Précis of *The Emperor's New Mind: Concerning computers, minds, and the laws of physics*

Roger Penrose
University of Oxford, Mathematical Institute, 24-29 St. Giles, Oxford, England OX1 3LB

*Abstract can be found on page 705.

In *The Emperor's New Mind* (1989) [henceforth *Emperor*] I attempt to put forward a point of view (which I believe to be new) concerning the nature of the physics that might underlie conscious thought processes. As part of my argument, I point out that there could well be room, within physical laws, for an action that is *not algorithmic* – i.e., that cannot be properly simulated by any computer – though I argue that it is likely that such nonalgorithmic action can arise only in an area of physics where there is an important gap in our present physical understanding: the no-man's-land between quantum and classical physics. (*Mathematical* processes of a nonalgorithmic kind certainly do exist, but the question I am raising is whether such processes have a role to play in *physics*.) I also argue that there is good evidence that conscious thinking is itself not an algorithmic activity, and that consequently the brain must be making use of nonalgorithmic physical processes in an essential way whenever consciousness comes into play. There must accordingly be aspects of the brain's action that cannot be properly simulated by the action of a computer, in the sense that we understand the term "computer" today.

Thus, the viewpoint I am putting forward dissents both from "strong-AI" (or "functionalism") – as expounded by Minsky (1968) – and also from a frequently argued contrary viewpoint, promoted particularly by Searle (1980). Strong-AI asserts that the brain's action is just that of a computer, conscious perceptions arising as manifestations of the mere carrying out of computations; this contrary viewpoint asserts that although computation does not in itself evoke consciousness, a simulation of the action of the brain would nevertheless be possible in principle, since the brain is a physical system behaving precisely according to some well-defined mathematical action. My dissent from this contrary view stems from the fact that "well-defined mathematical" does not, in itself, imply "computable."

Emperor's scope is broad, for it is my belief that there are many seemingly disparate topics that could have profound relevance to this important question. Moreover, I believe that real progress cannot be made into the deep philosophical issues raised by the question of "mind" without a genuine appreciation of the physical (and mathematical) principles underlying the actual behaviour of the universe in which we find ourselves. I have therefore tried to write this book at a level which makes it accessible, at least in principle, to readers without prior knowledge of the diverse topics covered. These topics include: the basics of artificial intelligence; Turing machines, computability and noncomputability; the Mandelbrot set and the system of complex numbers as illustrations of the Platonic world of mathematics; the foundations of mathematics, Gödel's theorem, nonalgorithmic mathematics; complexity theory; overviews of classical physics (including the issues of computability, determinism, and "chaos" within the theories of Newton, Maxwell, and Einstein) and quantum physics (its basic structure and its puzzles and paradoxes, the two types of evolution **U** and **R**); the second law of thermodynamics and its relation to cosmology, the big bang, black (and white) holes, the "improbability" of the universe, and the Weyl curvature hypothesis; the challenge of quantum gravity and its suggested role in the resolution of the puzzles of quantum theory; structure of the brain and nerve transmission; possible classical and quantum computer models; consciousness and natural selection; the nonalgorithmic nature of mathematical insight, the nature of inspiration, nonverbal thinking; the anthropic principle; a suggested analogy between nonlocal (quasi)crystal growth and the continual changes in brain structure, providing a possible input for (nonalgorithmic?) physics at the quantum-classical borderline; the singular relation between time and conscious perception.

Most of the book is noncontroversial and is intended to provide the reader with an overview of all the relevant topics needed. Though prior knowledge is not assumed, the presentation is of a sufficient depth that some genuine understanding of this material can be obtained. This is not always an easy matter, and parts of the book would need to be studied at some length if the arguments are to be fully grasped. There are places where I present viewpoints that deviate markedly from what might be considered to be "accepted wisdom." I have always been careful to warn the reader whenever I am presenting such unconventional views, even though I may believe the reasons pointing to the necessity of such "unconventionality" to be compelling.

The central questions we must ask are: Are minds subject to the laws of physics? What, indeed, are the laws of physics? Are the laws of physics computable (i.e., algorithmic)? Are (conscious) thought processes computable?

In my own view, mental states would be qualities that indeed depend on those same physical laws that govern inanimate objects. The minds we know of are features of the activity of brains (human brains, at least, but quite probably certain animal brains also) and human brains are part of the physical world. Thus, the study of mind cannot be divorced from the study of physics. Does this mean that new physical understanding is needed, or do we already have sufficient knowledge of the physical laws that might be relevant to an understanding of mental phenomena? Apparently, in the opinion of most philosophers, physicists, psychologists, and neuroscientists, we already know all the physics that might have relevance to these issues. I shall beg to differ.

It would, of course, be generally admitted by physicists that there are still many gaps in our understanding of the principles that underlie the behaviour of our universe. We do not know the laws that govern the mass-values or the strengths of the interactions for the menagerie of fundamental particles that are known to exist. We do not know how to make quantum theory fully consistent with Einstein's special theory of relativity, and certainly not with his general relativity theory. As a consequence of the latter, we do not understand the nature of space at a scale 10^{-20} times smaller than the dimension of the known subatomic particles, though our knowledge is presumed adequate at larger scales. We do not know whether the universe is finite or infinite, either spatially or temporally. We do not understand the physics that must operate at the cores of black holes or at the big-bang origin of the universe itself. Yet all these issues seem quite remote from what is relevant for understanding the workings of a human brain.

I shall argue, nevertheless, that many of these matters are of some relevance and, moreover, that there is *another* vast unknown in our physical understanding at just such a level as could indeed be important for the operation of human thought and consciousness – just in front of (or rather behind) our very noses! I believe there is a fundamental gap in our physical understanding between the small-scale "quantum level" (which includes the behaviour of molecules, atoms, and subatomic particles) and the larger "classical level" (of macroscopic objects, such as cricket balls or baseballs). That such a gap exists is a matter of physics – unrelated, at least in the first instance, to the question of "minds." I argue the case that there is good reason, on purely physical grounds, to believe that some fundamentally new understanding is indeed needed here – and I make some suggestions concerning the area of physics (the "collapse of the wave function" as an objective "quantum gravity" effect) wherein I believe this new understanding is to be sought. This whole issue is a matter of dispute amongst physicists at the moment; and it would have to be admitted that the apparent majority view is that no new theory is needed (although such outstanding figures as Einstein, Schrödinger, de Broglie, Dirac, Bohm, and Bell have all expressed the need for a new theory). It is my own strong belief that a radical new theory is indeed needed, and I am suggesting, moreover, that this theory, when it is found, will be of an essentially noncomputational character.

I argue that some of the manifestations of consciousness are demonstrably nonalgorithmic, and I am therefore proposing that conscious mental phenomena must actually depend upon such noncomputational physics. Although I try to make a fairly specific suggestion as to where in brain action the role of the physics governing the no-man's-land between quantum and classical physics might lie, my arguments are not critically dependent upon this particular suggestion. There might well be other possible roles for such (nonalgorithmic) physics in brain action. According to my viewpoint, the outward manifestations of the phenomenon of consciousness could never even be properly *simulated* by mere computation. Any properly "intelligent machine" – and I would argue that for it to be properly "intelligent" it is necessary for it to be conscious – could not be a "computer" in the sense in which we understand that term today (and so, in my own terminology, I prefer not to call such a putative object a "machine" at all), but would itself have to be able to harness the very nonalgorithmic physics that I am arguing must be a necessary ingredient of the physical basis of conscious thought. At present we totally lack the physical understanding to be able to construct such a putative "machine," even in principle.

Chapter 1: Can a computer have a mind?

Recall the Turing test, according to which a computer, together with some human volunteer, are both hidden from the view of some (perceptive) interrogator. The interrogator has to try to decide which of the two is the human being and which is the computer merely by putting probing questions to each of them, the answers being all transmitted in an impersonal fashion: say typed on a keyboard and displayed on a screen. The human subject answers the questions truthfully and tries to persuade the interrogator that he is indeed the human being, but the computer is allowed to "lie" so as to try to convince the interrogator that it, instead, is the human being. If on a series of such tests the interrogator is unable to identify the real human subject in any consistent way then the computer (or the computer's program) is deemed to have passed the test.

Setting aside, for the moment, the issue of whether, or when, some computer might actually pass the Turing test, let us suppose, for the sake of argument, that such machines have already been constructed. We must address the operationalist's claim that such a computer must be said to think, feel, understand, *et cetera,* merely by virtue of its passing. It is certainty my own view that such mental qualities – and certainly the central one of consciousness – are objective physical attributes that an entity may or may not possess. In our Turing-test probing, we are merely doing our best to ascertain, using the only means available to us, whether the entity in question (in this case the computer) has the physical attributes under consideration (in this case, say, consciousness). To my mind, the situation is not different in principle from, for example, an astronomer trying to ascertain the mass of a distant star. Being able to give human-like answers to Turing-test questioning is certainly not the same as having human-type mental qualities, but it may serve as a good indication that such mental qualities are indeed present in the machine. This would be in the absence of other criteria which might have a greater scientific reliability. In the case of mental phenomena, we do not really have better criteria than "communication" in some form, as with the Turing test. If someday we have a better theory of consciousness, then better criteria might eventually become available, and the Turing test will lose some of its importance. (Compare this with the possibility of accurately observing a planet in close orbit about our distant star. Einstein's theory can then provide a definite measurement of the star's mass.) In the absence of such a theory, however, the case for regarding the successful passing of a Turing test as a valid indication of the presence of thought, intelligence, understanding, or consciousness, and so forth, is fairly strong. For conversation is how we normally form our judgements that people other than ourselves actually do possess consciousness.

How long will it be before a computer actually passes the Turing test? It really depends on how strict one's criteria are for passing. It is my own guess that a proper passing of the test could not be achieved at all by an algorithmic computer (i.e., by a computer based on the calculational principles that we use today) – at least not in the foreseeable future. I have been vague about what constitutes a "proper" passing of the test, but extensive to-and-fro probing by the interrogator would certainly be required. (Turing originally suggested that a 30% success rate for the computer, with an "average" interrogator and just five minutes questioning, might be achieved by the year 2000.) In some clearly delineated fields, however, very impressive "human-like" behaviour has already been achieved. Chess-playing computers provide good examples of machines exhibiting what can be thought of as "intelligent behaviour." "Deep Thought" (programmed largely by Hsiung Hsu) has achieved some notable victories in some games with grandmasters. I think that it is clear, however, that computers play chess very differently from the way human beings do; computers relying much more on depth of extensive calculation and much less on "intuitive judgements" – whatever *they* are!

Perhaps a little more directly along the lines of a Turing test would be the computer programs of Schank and Abelson, as cited by John Searle (1980) in connection with his Chinese room argument. The computer using such a program is able to answer simple questions about stories (concerning, say, people eating in restaurants) in a way that a human being might. This seems to suggest that there is some primitive understanding on the part of the computer, since a human being would require some understanding in order to answer the questions correctly. Searle argues that no actual understanding can be taking place in the computer and that this limited "passing" of a Turing test is accordingly no indication that the mental quality of "understanding" actually occurs. To make his case, Searle envisages a "gedanken experiment" whereby the stories (about restaurants), the questions about these stories, and the computer's answers are all written in Chinese. The actual rules the computer follows in forming its replies to these questions are then laboriously followed by Searle himself (in a locked room), the rules for manipulating the relevant Chinese characters according to the computer's program being presented to Searle in English. Though Searle sits in this room and follows all the relevant actions of the computer, finally providing the correct (Chinese) answers to the questions, he does not himself obtain the necessary understanding of what the stories are actually about, since he knows nothing of the Chinese language.

Searle's argument is directed against the standpoint of "strong AI" (or functionalism) which claims that it is merely the enaction of a (sufficiently elaborate) algorithm that evokes mental qualities such as consciousness, understanding, or intentionality. I regard Searle's argument as quite persuasive with regard to programs of the fairly limited complication of the ones envisaged above, but it is by no means conclusive – especially when it is applied to the immensely more complicated putative computer programs that might, according to the strong-AI view, be necessary to conjure up actual consciousness! Perhaps there is some kind of "critical" amount of complication necessary for consciousness to be genuinely conjured up? Moreover, it is not a logical necessity that an algorithm's putative "awareness" should necessarily impinge upon the awareness of a person carrying out the algorithm, irrespective of the complication of that algorithm.

Searle is prepared to accept that a computer simulation of the actual activity of a human brain – when it is in the process of evoking conscious mental perceptions – would be possible in principle. But he believes that such a simulation would not, in itself, evoke the same mental perceptions. That certainly seems to me to be a tenable position, but not a very helpful one for gaining any understanding of why some objects enacting algorithms (brains) seem to evoke such mental qualities whereas others (e.g., electronic computers) are not supposed to. In later chapters, I present a different kind of argument to suggest that conscious mental activity is not algorithmic at all, and such a simulation would therefore *not* be possible.

The viewpoint of strong AI appears to spring partly from the fact that a person's individuality does not depend on the particular atoms that compose his body. There is certainly a continual turnover in virtually all the material of any living person's body. Moreover, according to quantum mechanics any two particles of any particular kind are completely identical to one another. For example, if the entire material content of a person were exchanged with corresponding particles in the bricks of his house then nothing would have happened whatsoever. What appears to distinguish the person from his house is the pattern in which his constituents are arranged, not the individuality of the constituents themselves. Thus, a person is just a "pattern of information," and this information could, in principle, be translated from one material form into another. The strong-AI viewpoint is that it is simply the "information content" of this pattern that characterizes any particular individual. This idea has gained strength from experience with modern high-speed computers, where we are now very familiar with the phenomenon of information being transformed from one realization into another (say from a pattern of magnetic fields in a floppy disc to a collection of charge displacements in a computer memory, and from there to a family of gaps in the illumination of a cathode ray screen). Moreover, there is the theoretical justification from the fact that modern general purpose computers are all, in effect, examples of *universal Turing machines* (see next section). Any two such machines are completely equivalent to one another in the sense that, irrespective of the particular hardware that may be involved in each machine, there will always be some appropriate software that effectively converts either machine into the other. In this sense, the hardware is regarded as being "irrelevant," and the essential "information" of the operation of the machine is considered to lie in its program, that is, software.

In mathematical terms, the program is effectively what is called an algorithm, that is, a step-by-step mechanical procedure for working towards an answer from any given input data. According to the strong-AI viewpoint, it is the very enacting of an appropriate algorithm that can (or must?) evoke (or be?) conscious awareness. Whether or not one believes in its essential correctness, however, there is much that is unclear about this picture. What does it actually mean to implement an algorithm, or to embody it in physical form? Do all the individual steps have to be performed in full by moving about bits of matter, or might a static printed description of the procedures be sufficient in itself to evoke conscious awareness? If the nature of the material performing the algorithm is irrelevant, then why is it necessary to have the material there at all? Should the algorithm's very (timeless!) Platonic existence not be sufficient to evoke consciousness? If it is, where does the distinctly temporal quality of consciousness come from? None of these matters seem to me to have been properly addressed by the proponents of strong AI.

An idea frequently discussed in this context is the teleportation machine of science fiction. A would-be traveller is scanned from head to toe, the accurate location and complete specification of every atomic nucleus and every electron in the subject's body being recorded in full detail. All this information is then beamed (at the speed of light), by an electromagnetic signal, to the distant planet of intended destination. There, the information is collected and used as the instructions to assemble a precise duplicate of the traveller – together with all his memories, his intentions, his hopes, and his deepest feelings, where the original copy of the traveller is to be destroyed in the process. Could a teleportation machine work – in the sense that the traveller's actual conscious identity is transferred to the distant planet? If teleportation is not travelling, then what is the difference in principle between it and just walking from one room into another? In the latter case, the traveller's atoms of one moment would simply be providing the information for the locations of his atoms of the next moment, and we have seen, after all, that there is no significance in preserving the identity of any particular atom. The moving pattern of atoms simply constitutes a kind of wave of information propagating from one place to another, so where is the essential difference between this and the teleportation device?

I believe that despite the outlandish nature of the teleportation idea, there is perhaps something of significance concerning the physical nature of consciousness and individuality to be gained from it. It may provide one pointer, indicating a certain essential role for *quantum mechanics* in the understanding of mental phenomena, for it turns out to be impossible to copy a quantum state unless the information in the original state is destroyed. If a person's individual awareness depends on some essential aspect of a quantum state, then it would not be possible to "teleport" that awareness unless the original copy of the individual were indeed destroyed. Accordingly, teleportation would not be impossible in principle; but these considerations would rule out the more paradoxical possibility of having two or more viable conscious copies of one individual.

Chapter 2: Algorithms and Turing machines

One of the most important developments in the history of mathematics occurred in the late 1920s and 1930s, when the concept of a general algorithm was made mathematically precise and it was demonstrated that there are some mathematical procedures that cannot be described by any algorithm whatsoever. There are various different but completely equivalent ways of formalizing the algorithm concept. The one which is intuitively the clearest was first put forward in 1935 by Alan Turing, called a *Turing machine:* an idealized device in that it operates with a potentially infinite "tape" on which the input data and instructions are represented. Only a finite portion of the tape is to be marked with actual data or instructions, however. Thus, although the device can cope with an input of unlimited size, this input is always finite. We may take the input and output to be recorded on the same tape; all the input goes in on the right and the output finally goes out on the left. For simplicity, we regard the tape as marked just with the symbols 0 and 1; "0" stands for blank tape and "1" for a mark on the tape (any more complicated symbols we might wish to use are broken down into a sequence of "marks" and "blanks"). The device itself has a finite number of distinct "internal states" – coded, for convenience, by binary numbers 0, 1, 10, 11, 100, 101, etc. – and it reads the symbols on the tape one at a time. It has a finite list of instructions that tell it what to do, given its internal state and whether it reads 0 or 1. The list of instructions itself is never altered, but the particular 0 or 1 that is being read by the device may either be left alone or else changed to the opposite symbol, depending upon the particular instruction that comes into play; the device also accordingly moves one step to the right, one step to the left, or comes to a halt after moving one step to the right. (In these descriptions I am taking the device to be moving rather than the tape, but the two pictures would be equivalent.) As part of the same operation, the internal state is changed to another internal state, as specified by the instruction. Now, in this new internal state, the device (if it has not come to a halt) reads a new 0 or 1 on the tape and acts according to the appropriate instruction that now comes into play. The process continues until a halt instruction is encountered; at that point we imagine a bell ringing to alert the operator of the machine that the calculation has been completed.

It is a remarkable fact that any computational process whatever (that operates with finite discrete quantities) can be described as the action of some Turing machine.

This, at least, is the contention of the so-called Church-Turing thesis, in its original mathematical form. Support for this thesis comes partly from Turing's careful analysis of the kinds of operation one would actually consider as constituting a computational or algorithmic process, and partly from the striking fact that all the various alternative proposals for what an "algorithm" should mean (put forward at around the same time by Church, Kleene, Gödel, Post, and others) have turned out to be completely equivalent to one another. Some of these proposals had the initial appearance of being completely different, so their equivalence is a strong indication of the fact that they are merely alternative ways of describing an absolute abstract mathematical concept, that of *computability*, (which is independent of any particular realization of it that one may care to adopt. (In addition to an extended and detailed description of Turing machines, I give a brief description of Church's remarkable calculus in *Emperor*, pp. 66–70).

Like so many other mathematical ideas, especially the more profoundly beautiful and fundamental ones, the idea of computability seems to have a kind of *Platonic reality* of its own. This mysterious question of the Platonic reality of mathematical concepts is a central theme of much of *Emperor*, and I shall need to return to it later.

Having a very specific description of what a Turing machine is, we may turn to the *universal* Turing machine. This is a particular Turing machine U (having one fixed list of instructions) which can mimic any other Turing machine. All that is needed is that the Turing machines be coded in a well-defined way, so that each machine is assigned a number unique to that particular machine. The Turing machine with number n is denoted by T_n, and is referred to as "the nth Turing machine." To make U act like the machine T_n, all we need do is feed U with the number n on the tape first (say coded in the binary notation) and then (on the right of a suitable coded marker) we would feed in the tape that the machine T_n is supposed to be reading. The final result of the action of U on the whole combined tape would be the same as that of T_n on the right-hand part alone. Note that the machine U itself must have a number. In *Emperor* (pp. 56–57; 71–73) I exhibit this number (or one such – there are many possibilities) explicitly. It turns out to be about 7×10^{1654}.

Turing originally devised his "machines" to answer a question posed by David Hilbert: Is it possible in principle to find a mechanical mathematical procedure (i.e., an algorithm) for answering all mathematical problems in a specified class? Turing (and, independently, Church) showed that the answer is "No." Turing phrased his version of Hilbert's problem as the question of deciding whether or not a given Turing machine, when acting on a specific tape, will ever come to a halt; he showed (using a "diagonal argument") that there is *no* algorithm for systematically answering this question.

Many people seem to be under the impression that this means there are specific Turing-machine actions for which it is impossible to decide whether or not they halt. This is not what the argument shows at all, however. It merely shows that there is no *algorithmic procedure* for deciding this question *in general*. In any specific case, there is certainly such an algorithm, namely, the algorithm that simply says "yes" or else the algorithm that simply says "no" – but, of course, we would not know which of these two algorithms is the correct one to use! Algorithms do not in themselves decide questions of truth or falsity. For that we require *insights*, not algorithms. To illustrate this point, I give a modified version of Turing's original argument. I show that if we are given any algorithm that correctly decides whether Turing machine actions halt, though for some such actions that algorithm may itself run on forever without stopping, then we can exhibit a specific Turing-machine action that *we* can see does not come to a halt but for which the given algorithm never comes to a decision. In this sense we can in principle, by the use of insight, "outdo" any given algorithm for testing whether or not Turing machine actions halt.

Chapter 3: Mathematics and reality

The Mandelbrot set provides a wonderful illustration of a mathematical structure that, though defined in an entirely abstract mathematical way, nevertheless has a reality about it that seems to go beyond any particular mathematician's conceptions and beyond the particular technology of any specific computer. Though more and more of its wonderfully elaborate structure is revealed to us as more computer power is brought to bear on the problem of exhibiting the set, there is always more of the set to be found that is still hidden from us. The set seems clearly to be "there," somewhere, quite independently of us or of our machines. Its existence is not material, in any ordinary sense, and it has no spatial or temporal location. It exists, instead, in Plato's world of mathematical entities. When we use our computers to explore the set, it is like using a moon-rover to explore the moon's surface or a high-energy particle accelerator to probe the secrets of subatomic physics.

The precise mathematical definition of the Mandelbrot set is remarkably simple, considering the extraordinary complexity of its detailed shape. But to understand this definition, it is first necessary to come to terms with the idea of a complex number. It turns out that this is just as well in any case, for complex numbers are absolutely fundamental also to the structure of quantum theory, a theory that we need to come to terms with later.

Complex numbers are numbers of the form $x + iy$, where x and y are ordinary "real" numbers (i.e., numbers that can be expressed in terms of infinite decimal expansions) and where "i" represents a square root of -1. In ordinary terms, one tends to think that -1 does not "really" have a square root, and that it is just an "invention" on the part of mathematicians to suppose that it does. But the so-called "real" numbers are taken as more real only because we have got much more used to them, and because they accord remarkably closely with physical measurements such as time, distance, energy, temperature, and the like. There are reasons to expect that even this close correspondence with physical quantities may break down at a very tiny scale, and we must accept that the utility of the concept of a real number lies as much in its mathematical consistency, power, and elegance as it does in any correspondence with the physical world. We find, indeed, that the concept of a complex number leads us to a picture with perhaps an even greater power and elegance than that of a real number; moreover, complex numbers also have a profound relationship with the workings of the physical world.

The very system of complex numbers provides further evidence for a Platonic existence for mathematical entities. While complex numbers were first introduced as a kind of "trick," useful in the solution of cubic equations, they were later found to have enormous power and utility in many different areas. In general, the case for Platonic existence is strongest for entities which give us a great deal more than we originally bargained for. In such cases, most of what we get out may be something that was not even remotely conceived by those mathematicians who first came across the entities in question. The mathematicians discovered something wonderful. They did not invent it. The Mandelbrot set, and even more so, the very system of complex numbers, are clear cases in point. There are many other situations in which the case for Platonic existence is not so strong, however, such as when a mathematician introduces a construction of no particular elegance or uniqueness to prove a particular result, this construction finding no value other than the one for which it was originally devised. In such cases it is not at all unreasonable to use the word invention, rather than discovery. It is the clear discoveries in mathematics that are the mathematician's proudest achievements.

The complex numbers have an elegant geometrical representation in the Euclidean plane, where the number $z = x + iy$ is plotted as the point with coordinates (x,y). A plane describing complex numbers in this way is called an *Argand plane*. The Mandelbrot set is a subregion of the Argand plane defined in terms of iterations of the map whereby z is replaced by $z^2 + c$, starting from $z = 0$, c being some fixed complex number. If these iterations lead to an unbounded sequence of points in the Argand plane then the point representing c lies outside the Mandelbrot set. If it is a bounded sequence, then c is in the Mandelbrot set.

One final comment about real and complex numbers should be made here. There are, in a precise sense, many "more" of them than there are natural numbers 0, 1, 2, 3, 4, 5, 6, 7, . . . to which the usual discussions of Turing machines refer. We might think of using some algorithm for generating the successive digits of an infinite decimal expansion, but it turns out that only a tiny fraction of the possible decimal expansions are obtainable in this way: the *computable numbers*. (Nevertheless, all the familiar numbers, such as the decimal expansion of pi, are computable.) This causes some problems when one tries to apply the concept of computability to descriptions of the physical world, which normally use real or complex numbers (however, see Blum et al. 1989).

Chapter 4: Truth, proof, and insight

The question of mathematical *truth* and how we ascertain it is fundamental to our considerations here. How do mathematicians actually decide which mathematical statements are true and which are false? Are they following some algorithm – perhaps one that is unknown or unknowable to them – or do they have some other route to truth via some mysterious "insight" that is not amenable to a purely algorithmic treatment? To gain some understanding of this issue, it is necessary to go back a little into the mathematical history of the last century or so. Mainly as a result of work by the mathematician Georg Cantor, it was found that very powerful methods of reasoning can be obtained if one is prepared to accept infinite sets as entities that can have independent existence in their entirety. It was soon found (by Bertrand Russell and others), however, that contradictory results are liable to arise with this kind of reasoning unless some special rules are introduced to limit the "sizes" of the sets under consideration. What should these rules be? Russell, Hilbert, and various other mathematicians proposed some very precise systems of axioms and rules of procedure – called formal systems – with the intention that these should incorporate all the legitimate rules of mathematical procedure. If they were to be completely successful, then any true mathematical proposition (within some well-defined area of mathematics) ought to be provable by means of the procedures laid down in the formal system (completeness) and, moreover, it should not be possible to prove both a proposition and its negation (consistency). This led to the mathematical standpoint of *formalism*, according to which mathematics could be reduced to a kind of "game" consisting of merely manipulating symbols according to the specific rules of the formal system in question. The necessity for the symbols to have any actual "meaning" would thereby be eliminated.

Fortunately (to my way of thinking), in 1931, Kurt Gödel presented his famous theorem, which effectively destroyed formalism as a fundamental philosophy of mathematics. He showed that in any consistent formal system that is broad enough to contain arithmetic and the normal rules of logical procedure one can explicitly construct well-defined mathematical statements that are not provable – nor are their negations provable – using the rules laid down in that formal system. Thus, the system cannot be complete, in the sense that Hilbert and others required. Worse than this (from the formalist's point of view) by the very way that such a Gödel proposition is constructed we can *see*, using our insight and understand about what the symbols in the formal system are supposed to mean, that the Gödel proposition is actually *true!* This tells us that the very concepts of truth, meaning, and mathematical insight cannot be encapsulated within any formalist scheme.

This is not just bad news for the formalists. It is bad news for strong-AI, as well. For there is a very close relationship between the concept of an algorithm and the concept of a formal system, with regard to mathematical statements. For any formal system, there is always an algorithm that generates precisely all the propositions that can be proved within that system. Conversely, given an algorithm for generating mathematical statements, one can always construct a formal system that incorporates all these statements as axioms of the system. This tells us that mathematical truth is not an algorithmic matter; It also appears to tell us that meaning and insight are not algorithmic matters either. I return to this issue in Chapter 10.

Since Gödel's theorem seems to be telling us that formalism is not tenable as a foundation for mathematics, are we to be driven to a *Platonic* view that mathematical concepts are in some sense just "out there" waiting to be discovered? In my own view, something like this must be right, though many people find it hard to believe. In particular, there is another mathematical standpoint, referred to as ***intuitionism***, according to which completed infinities are not permitted, and are allowed only to have a potential existence. (The intuitionistic philosophy, due mainly to L. E. J. Brouwer, stems more from the ideas of Aristotle than those of Plato.) In this view, what is important is that abstract proofs of "existence" are not allowed unless an explicit construction of the quantity in question is given. The main problem with intuitionism lies in the severe limitations that it places on mathematical reasoning. (In particular, the powerful method of ***reductio ad absurdum***, which is used so frequently in mathematics, is not now permitted.) Moreover, at least in the way it is often presented, the intuitionistic concept of truth has a subjectivity and time-dependence that seems very unsatisfactory for a mathematics that one hopes to be able to use reliably in a description of the physical world.

Accepting that much of mathematics is of a nonalgorithmic character, we can ask whether nonalgorithmic mathematics is in any way interesting or mathematically important. In fact, it is possible to give numerous simple and interesting examples of nonalgorithmic mathematics. Diophantine arithmetic (the question of whether systems of polynomial equations in integers, with several variables, have solutions) turns out to be nonalgorithmic (this was actually Hilbert's original question); so does the topological equivalence problem for 4-manifolds (of possible relevance for the problem of quantum gravity) and, perhaps most striking, the problem of deciding whether or not a given finite set of polygonal shapes will tile the entire Euclidean plane. In each case, we have a family of well-defined yes/no questions for which it is known that no algorithmic solution exists. In *Emperor* I also conjectured that the problem of deciding whether or not a point of the Argand plane actually lies in the Mandelbrot set is, in an appropriate sense, nonalgorithmic[1] – despite the fact that computers are able to provide wonderful approximations to this set!

Another body of understanding that has considerable importance for computability problems is *complexity theory*. As with questions of algorithmic solubility, complexity theory is concerned with families of mathematical problems, but where the problems in each family are taken to be algorithmically *soluble*. The problem is to decide how "good," in some well-defined sense, such an algorithm can be for a given family of problems. It turns out that for some there are fast algorithms (the so-called *P* problems), while for others the algorithms seem to be of necessity so slow that the problems are, in effect, intractable (i.e., insoluble algorithmically "in practice" – e.g., the so-called NP-problems that are not in *P*). In my opinion, the questions of complexity theory are not so central to the issue of conscious thinking as are the questions of computability theory, but many people seem to hold to the contrary. [See Tsotsos: "Analyzing Vision at the Complexity Level" *BBS* 13(3) 1990.]

Chapter 5: The classical world

We now turn to the question of how the physical world actually behaves. The issue of *determinism* in connection with physical laws has often been discussed. I am more concerned with a different (though somewhat related) issue that, with the important exception of some significant work by Pour-El and Richards (1981; 1982), has barely been addressed at all: that of *computability* in physical laws.

Before we can come properly to terms with these issues, however, we must have some idea of what the laws of physics actually *are*. It is a very remarkable fact that there are indeed physical laws that have a truly *phenomenal* accuracy and, moreover, are extraordinarily amenable to precise and general mathematical treatments. These physical laws provide the theories that I refer to as SUPERB. These theories are supplemented by others that help to explain physical phenomena but do not share with the SUPERB theories their phenomenal accuracy. Such theories I refer to as USEFUL. In addition, there are numerous ideas, some currently very fashionable, and often excitedly expressed, that (essentially) have no experimental support whatever. These are theories belonging to the class TENTATIVE. In *Emperor* I give a brief account of each of the SUPERB theories in turn.

The oldest of the SUPERB physical theories is Euclidean geometry, which provides a marvellously accurate theory of physical space and of the behaviour of rigid bodies. The fact that Euclidean geometry is an extraordinarily precise physical theory, and not just a very elegant area of pure mathematics, is demonstrated (perhaps somewhat ironically) by the fact that we now know that this geometry is actually not exactly true of the physical space we inhabit. The even more accurate geometry of Einstein's general relativity provides a picture of a curved space(-time) that deviates in very tiny and barely measurable ways from the geometry of Euclid. To qualify as SUPERB, it is not necessary that the theory in question apply without exception to the observed properties of the world, only that its agreement with those properties qualify as "phenomenal" in some appropriate sense. Be that as it may, there is one essential feature of Euclidean geometry (introduced by Eudoxos in the fourth century B.C.) that remains with us to this day as a fundamental ingredient of *all* the SUPERB (and also USEFUL) theories: the concept of a *real number*.

Probably Archimedes's theory of statics would have qualified as SUPERB, but this theory is now subsumed into the SUPERB dynamical theory of Galilei/Newton. *Newtonian mechanics* is, as we all know, deterministic, but is it *computable*? Suppose initial data for some physical situation is given in terms of computable numbers (all constants involved being also computable numbers), and we wait for a computable time. Is the state at that time computable from the initial data? As far as I am aware, the question has not yet been properly addressed mathematically, though my guess is that the answer would be "yes" in Newtonian theory (with reasonable force laws – say inverse square, with hard elastic spheres and "generic" initial data, so as to avoid the problem of triple collisions).

It should be made clear that the possible "noncomputability" that is being considered here is different from the much-discussed concept of "chaos." *Chaos* refers to the fact that very tiny changes in initial data may give rise to absolutely enormous changes in the resulting behaviour. Chaotic systems would normally be computable, in my sense. (The issue of chaos is like that of complexity theory, as discussed above, rather than computability theory.) Chaotic systems are noncomputable *in practice*, rather than *in principle*. In practice, they introduce a random element into the evolution. Such a random element is not the kind of thing that could be usefully "harnessed" by the brain in order to achieve something nonalgorithmic. Newtonian mechanics can be incorporated into the more general framework of *Hamiltonian mechanics*, which can also be made to include Maxwell theory and relativity. In fact, however, most Hamiltonian systems appear to be chaotic. It seems likely to me, however, that normal Hamiltonian systems (with computable constants) should be *computable*.

Maxwell's theory provides the SUPERB equations governing electric and magnetic fields, and it shows how these fields propagate through space in the form of light or radio waves. The radical new ingredient that Maxwell (and Faraday) introduced was a physical entity with an infinite number of degrees of freedom. Like Newton's laws, Maxwell's equations are deterministic and presumably computable in some appropriate sense. (Some curious problems in relation to this have been pointed out by Pour-El and Richards [1981], although it seems likely that the "noncomputability" they encounter is not of physical importance, arising, as it does, only with data that are not smooth.) To describe the motion of charged particles in addition to electromagnetic fields, Maxwell's equations have to be supplemented by the *Lorentz force law*. The Maxwell-Lorentz equations work well in practice but can lead to problems of principle with regard to determinism, as was pointed out by Dirac. This anomaly is usually ignored on the grounds that one should really be using quantum theory for such situations (though the quantum theory is actually no better!).

Einstein's SUPERB *special theory of relativity* changes our picture of space-time, but it does not seem to alter the issues of determinism and computability (except slightly to improve the situation with regard to determinism). It does radically affect the way that we view the "passage of time," however. His even more SUPERB *general* theory of relativity asserts that space-time is curved and that this curvature describes the gravitational field. An important property of space-time curvature is that it can be naturally split into two pieces, which I refer to as RICCI and WEYL. Einstein's field equations assert that RICCI is directly determined by the mass-energy distribution of matter (counting the electromagnetic field also as matter), and the remaining part, WEYL (which describes tidal distortion), represents the free gravitational field. It is conceivable that there are situations involving ultrastrong gravitational fields in which there is a failure of ordinary determinism (failure of "cosmic censorship"), but such situations could in any case occur only at a scale totally different from that of human brains.

Relativity provides us with a curious paradox concerning the nature of matter. Matter is quantified by its mass, but gravity itself has energy and therefore (by Einstein's $E = mc^2$) also mass. In certain circumstances, the mass of gravity appears to reside in regions of space-time that are not just empty but actually *flat!* I mention this merely to point out that even in classical theory there can be a puzzling nonlocality about the very nature of matter.

Such nonlocality is an even more puzzling with the mysterious but SUPERB theory of *quantum mechanics*, and also with that somewhat unsatisfying but powerful theory referred to as quantum field theory, which arises when the principles of special relativity are combined with those of quantum mechanics. The particular quantum field theory that applies to the electromagnetic field interacting with electrons (or with certain other particles such as protons) is called *quantum electrodynamics*, and it also qualifies as SUPERB. We consider quantum theory next.

Chapter 6: Quantum magic and quantum mystery

To have any hope of understanding the mysterious phenomenon of consciousness in physical terms, one must try seriously to come to terms with the physical laws that govern the way things *actually* behave in our universe. With the possible exception of the structure of space-time itself, the most fundamental and mysterious of these physical laws are the laws of *quantum mechanics*.

There is a basic but very puzzling quantum mechanical principle – called the *superposition principle* – that asserts that if A is a possible state of a system and if B is another possible state, then $wA + zB$ is also a possible state, where w and z are two complex numbers, not both zero (the ratio $w{:}z$ being what is physically important). What does this mean? In a particular instance, consider the state of a particle, which might be at one point (state A) or else at another (state B). The superposition principle tells us that another possible state is $A + B$ and another is $A - B$, while yet another is $A + iB$ where i is the square root of -1). Thus, not only must we consider that a particle can be "in two places at once," but also that there are many *different* ways of being at two places at once! All these different possibilities have different experimental consequences.

It appears that the superposition principle holds only at the "quantum level," where differences in alternative possibilities are in some appropriate sense small (small differences in energy distribution – or some such criterion), whereas at the "classical level," which seems to include the level of our ordinary experiences, the superposition principle seems to make no sense at all. Instead, we must adopt a new rule whenever the respective effects of A and B are in some way magnified to the classical level. This rule asserts that the so-called "probability amplitudes" – in effect, the complex numbers w and z in the above description – must have their *squared moduli* formed (their distances from the origin in the Argand plane), and the ratio $|w|^2{:}|z|^2$ provides the relative probabilities for each of the two alternatives in question *actually* to occur. I refer to this so-called "observation" or "measurement" process as **R** (*reduction* of the state vector, or "collapse of the wavefunction").

In my own opinion, **R** is a *real physical process* which takes place spontaneously in suitable physical circumstances, independently of human intervention or the introduction of "consciousness." Physicists seem to hold to various different views on this, many maintaining that **R** is in some way a kind of "illusion." Such theorists often find themselves driven to a very "subjective" view of physical reality at the quantum level. One reason physicists have trouble with **R** is that it is completely different from the process **U** that governs the behaviour of a state which remains at the quantum level. The action of **U** (*unitary* evolution, described by the important Schrödinger equation – or else by something equivalent such as the Heisenberg equation) is entirely deterministic, just like a classical field equation such as Maxwell's. The probabilities in quantum theory arise only with the process **R**, that is, when effects become magnified from the quantum to the classical level.

Some physicists (probably only a small minority) believe that the superposition principle should still be applied at the classical level (so **R** never actually takes place at all); they are led to the *many-worlds* (or Everett) interpretation of quantum mechanics (Everett 1957; DeWitt & Graham 1973), whereby all alternative possible universes must be considered to coexist in some vast complex-number superposition. No convincing explanation is provided, however, for why we (as conscious entities?) should perceive just *one* out of the infinitude of superposed alternatives, or why the quantum mechanical probability rules should apply.

Whether or not one adheres to this strange view, there seems to be no completely satisfactory alternative within the confines of standard quantum theory. The alternative viewpoints include the "subjective" one referred to above, according to which the quantum state (despite its SUPERB accuracy in providing agreement with observation) is not taken seriously as a description of reality ("all in the mind" of the experimenter!), and there is the alternative view that somehow the "illusion" of **R** occurs whenever the complicated interaction with the environment is taken into account. (I do not find either of these views, in themselves, to be very satisfying or convincing). Some people prefer the view (Wigner 1961) that a *conscious entity* (or perhaps just a "biological system") would not be subject to the superposition principle, and that consequently some *new nonlinear theory* is needed to handle the way quantum mechanics could be applied to such an entity. This would entail an actual change in the basic structure of quantum theory. I am personally very sympathetic to the general view that a fundamental change is needed, but I do not think that it is at the level of the phenomenon of consciousness (or biology). That would lead to a very lopsided view of the reality of the physical world. For in those corners of the universe in which consciousness (or biology) resides, the objective physical behaviour of matter would then be totally different from its behaviour everywhere else (where there would be planets with complex superpositions of many different weather patterns, etc. – and probably much worse!). Other deviations from standard linear quantum theory that do not involve consciousness and seem to me to have much greater plausibility (e.g., Ghirardi et al. 1986; Pearle 1989) have been proposed in order to resolve the very basic difficulties that arise from the **U**/**R** conflict. These difficulties are made particularly graphic in the famous Schrödinger (1935) "cat" thought experiment. I shall come to my own specific suggestions on the matter later.

One of the most puzzling aspects of quantum theory lies in the way that systems of *many particles* must be treated. (Complicated quantum systems do not behave like classical ones, just by virtue of their complication, despite what many physicists appear to think. Somehow **R** must come into it, but the matter remains very unclear.) When there are many particles, individual particles are not to be considered as objects on their own. Their quantum states involve superpositions of each particle's possible individual state combined with the various possible individual states of all the other particles in the system under consideration. (The different particles are then said to be *correlated*.) This has the effect that if one particle is "observed" (i.e., the particle triggers an effect that magnifies its situation from the quantum to the classical level), then this instantaneously affects the quantum state of all the other particles with which it is correlated. This effect leads to what is known as the *Einstein-Podolsky-Rosen* (EPR) *paradox* (Einstein et al. 1935). It is not really a physical paradox, but an actual physical effect that is not explicable on the basis of any local realistic view of physical behaviour. (This follows from a remarkable theorem by J. S. Bell (cf. Bell 1987) and is substantiated in the behaviour of the actual world in experiments such as those performed by A. Aspect et al. 1986).

In my own view, there is nothing in principle objectionable about a nonlocal (or somewhat "holistic") picture of physical reality, but there is a very serious difficulty with obtaining a picture that is consistent with the spirit (and space-time descriptions) of relativity. This presents a profound challenge for what I believe to be a much-needed new theory.

Chapter 7: Cosmology and the arrow of time

Why do we perceive time to "flow," when our (SUPERB) theories tell us that in "reality" there is just a static space-time laid out, with no "flowing" about it? Relativity even tells us that there cannot be any such things as "now" at all, since the very concept would depend on how various other observers might be moving. Even worse, apparently, is the fact that the equations of physics are *symmetrical in time*, so that they would apply just as well in the reverse temporal direction as in the normal forward direction. Why does time seem to "flow" forwards and not backwards?

At least a partial answer to this last question can be found in the *second law of thermodynamics*. This law is time-asymmetric; and it asserts that a certain physical quantity called *entropy* increases in the forward time-direction (and, consequently, it decreases in the backward time-direction). Very roughly speaking, entropy is a measure of the manifest disorder of a physical system. It is not unnatural physically that a system, when left on its own, should get more and more disordered. The *puzzle* of the second law lies in the fact that this is not what happens in the *reverse* direction in time or, to put things another way, that the state in the past was actually given to us as extraordinarily ordered. We can trace the nature of this "order" to the very structure of the *big bang* – the singular state (infinite space-time curvature) that, according to standard (USEFUL, at least) theory, represents the actual origin of the entire universe.

We can compare this initial singular state with the ones that are expected to occur in the reverse direction of time, at the cores of black holes – or in the "big crunch" that will finally engulf the entire universe, if our universe turns out actually to be finite rather than infinite. The geometrical difference, in the case of the big bang, appears to lie in the fact that in the immediate neighbourhood of this singular state the space-time curvature is constrained by the condition: WEYL = 0. The hypothesis whereby initial-type singular states (i.e., at the big bang, or in "white holes" – the time-reverses of black holes) are constrained by WEYL = 0, whereas those of final-type (i.e., at the big crunch, or in black holes) are not to be so constrained, is what I refer to as the *Weyl curvature hypothesis* (WCH). This hypothesis would explain both the puzzle of the second law and the observed extraordinary uniformity of the actual universe. Such a constraint could explain the fantastic geometric precision involved in the "act of creation" that occurred in the big bang – to one part in at least $10^{10^{123}}$ (a number so large that it could not remotely be written out in full even if each digit were to be written on each separate electron, proton, neutron, or any other particle in the entire universe!).

Chapter 8: In search of quantum gravity

What lies behind WCH? Could this hypothesis be a deduction from some comprehensive theory of physics that deals also with problems at other scales? The usual viewpoint about the singular states that classical general relativity leads us to (at the big bang, big crunch, and in black or white holes) is that they must be dealt with by a *quantized* general relativity, or *quantum gravity*. This is also my own view, but the fact that quantum gravity needs to explain a grossly time-asymmetric phenomenon like WCH does not seem to be generally appreciated. No satisfactory theory of quantum gravity has yet emerged, but when it does (and let us call this putative theory CQG – "correct quantum gravity"), it ought to turn out, on the basis of the above considerations, to be a time-*a*symmetric theory. This is very different from the kind of thing that has been expected until now, but considerations of gravity quantizers have so far been restricted to trying to derive a (time-symmetric) quantum procedure **U** for the time-symmetric classical theory of general relativity (or one of its equally time-symmetric generalizations). No one seems to have tried seriously to incorporate the time-*a*symmetric procedure **R** into quantum gravity theory. I give an argument in *Emperor* to show that **R** is indeed time-*a*symmetric (a fact not always recognized by physicists) and I describe a thought experiment ("Hawking's box") that strongly suggests that if CQG implies WCH, then CQG also ought to incorporate **R** as an actual physical process.

This leads to the suggestion that the onset of **R** can be understood as something that occurs when pairs of states are superposed and the differences in their gravitational fields (i.e., the differences in their space-time geometries) reaching the "one-graviton level," or thereabouts. Simple physical examples are described which indicate that this criterion is not at all physically unreasonable.

It seems to me that an important new physical theory must be lurking in the shadows. But recall the difficulties referred to above: Any realistic theory of **R** would have grave difficulties in coming to terms with the space-time descriptions that must accord with the principles of relativity. In my opinion, our present picture of physical reality is due for a grand shake-up – even greater, perhaps, than that which has already occurred with relativity and quantum theory.

Chapter 9: Real brains and model brains

What do we know of the actual structure and workings of the human brain? What aspect of the brain seems to relate most to the phenomenon of consciousness? There seems to be remarkably little consensus as to what actual parts of the brain are to be most associated with this phenomenon. The cerebral cortex, reticular formation, thalamus, hippocampus, and no doubt many other parts of the brain have each have been separately suggested as the place to look for the "seat of consciousness." Some people have suggested that only the left half of the brain is conscious, since (in most individuals) that is the half capable of speech – an odd view, to my mind, and certain experiments on split-brain patients indeed suggest that both halves can be conscious. Perhaps consciousness is not to be located in any one clear place. Some parts of the brain do seem to be more closely associated with consciousness than others, however. The cerebellum, for example, really does seem to act as an unconscious "automaton," whereas one does sometimes (although apparently only sometimes) seem to be aware of activity taking place in the cerebral cortex. This might seem surprising, on the basis of the strong-AI picture, the cerebellum having a much higher density of neurons than does the cerebral cortex, and perhaps half as many neurons altogether!

One can present computer models of the activity of the firing of neurons, but is this liable to give a reliable modelling of the brain's (conscious?) activity? There is at least one feature of brain functioning that is not well modelled in this way, and that is *brain plasticity*, according to which the connections between different neurons can become strengthened or weakened, providing the essential ingredient of permanent memory, according to one prevalent theory. There are suggested mechanisms underlying brain plasticity (such as the one due to Hebb 1954) and these can be modelled on computers as *neural networks*. The processes underlying *actual* brain plasticity seem to be largely unknown, as of now, however.

Parallel processing is another ingredient that seems to have importance in the computer modelling of brain function. It seems unlikely to me however, that any proper understanding of conscious thought processes will come about in this way. There is a remarkable "oneness" in conscious thinking that seems very much at odds with the functioning of a (classical) parallel computer. In any case, parallel computers are completely equivalent to the normal serial (Turing machine type) computers with regard to what they can and cannot compute.

Is there an essential role for quantum phenomena in brain functioning? Perhaps quantum superposition is somehow usefully incorporated into brain function. One might envisage that huge numbers of different calculations are carried out simultaneously in superposition, and only at the end is **R** called into play to conjure up the answer that is required. Such would be the action of a *quantum computer*, as considered by Deutsch (1985;

Deutsch's quantum computers, however, do not perform noncomputable operations – although in certain rather contrived situations they can do better than Turing machines in the sense of complexity theory.) It is known that there are neural cells (in the retina) that are sensitive to single quantum events (photon arrival), but it is hard to see how to harness this kind of thing in any useful way, the brain being too "hot" a system to preserve quantum coherence over an appreciable length of time.

Quantum considerations are relevant to many aspects of brain functioning (chemical transmission, the definite potential differences responsible for the on/off transmission of nerve signals, etc.), but it is difficult to see how present-day quantum mechanics could be coherently used to describe the action of a human brain, even in principle: For the brain would have to be considered as "observing itself" all the time – when it is conscious, at least, which would entail the continual use of **R**. There is no satisfactory theory for handling this.

Chapter 10: Where lies the physics of mind?

What is the selective advantage of consciousness? Our consciousness has presumably evolved because the behaviour of an animal with consciousness is actually more advantageous in some way to that of an otherwise equivalent animal without consciousness. If consciousness is merely the inevitable inward manifestation of the possession of a sufficiently complex control system, why is it that that some parts of the brain (e.g., the cerebellum) do not appear to be conscious, yet they can perform extremely complex tasks? What is advantageous about the outward manifestations of those particular thought modes that seem necessarily to be associated with consciousness?

If one contrasts the phrases (e.g., "common sense," "judgement of truth," "understanding," "artistic appraisal") that might be associated with many of those mental activities that require some degree of consciousness with those (e.g., "automatic," "following rules mindlessly," "programmed," "algorithmic") that do not seem to require it, then the possibility of a *non-algorithmic/algorithmic* distinction at least suggests itself. In my view this is one strong indication that what we are doing with our consciousnesses is actually not something algorithmic at all. As a central theme of *Emperor*, I have tried to stress that the mere fact that something may be scientifically describable in a precise way does not imply that it is computable. It is quite on the cards that the physical activity underlying our conscious thinking may be governed by precise but nonalgorithmic physical laws and our conscious thinking could indeed be the inward manifestation of some kind of nonalgorithmic physical activity. I am suggesting, therefore, that the selective advantage of consciousness is that it enables its possessor to form some kind of nonalgorithmic judgement of how to behave in a given situation.

Perhaps one could refer to such judgements as "inspired guesswork;" it is of some value to examine some outstanding examples of inspiration, as recorded by Poincaré and Mozart, for example. Here the inspirational ideas appear to be thrown up by the *un*conscious mind, but it is the *conscious judgements* that are needed to assess the value of the ideas themselves. These judgements have a remarkable globality about them; a vast area seems to be surveyable in an "instant" – say, an entire mathematical topic, or a symphony. Such globality is also a feature of much of our conscious thinking at a (seemingly) much more mundane level, such as deciding what to have for dinner or appreciating a visual scene.

Why do I maintain that such conscious judgements must have an essentially nonalgorithmic ingredient? One could certainly imagine feeding appropriate criteria into an ordinary algorithmic computer and getting it to produce "judgements." The most decisive reason for believing that our conscious judgements must be non-algorithmic comes from *mathematics*. If mathematical judgements can be seen to be nonalgorithmic – where the criteria of logic, precision, correct calculation, and truth are normally taken as paramount – then surely non-algorithmic ingredients could have at least as great an importance in other areas.

In Chapter 4 of *Emperor*, I show how to construct, for any (sufficiently broad) formal mathematical system, a specific Gödel proposition $P_k(k)$, which is a well-defined statement about numbers. It has the form "for every natural number x, the following computable property of x holds." From the way that $P_k(k)$ is constructed, one sees, provided that one believes that the axioms and rules of procedure of the formal system are valid methods of deriving mathematical truth, that one must believe that $P_k(k)$ is a mathematical truth *also*. Nevertheless, $P_k(k)$ is *not* itself derivable by means of the axioms and rules of procedure of the given formal system.

We recall that an essential equivalence exists between formal systems and algorithms as procedures for ascertaining the truth of mathematical propositions. Now suppose that a particular mathematician is using some algorithm – that is, in effect, some formal system F – as his means of ascertaining mathematical truth. Then the Gödel proposition $P_k(k)$ constructed from F must be a true proposition also, though it is not possible for our putative algorithmic mathematician to ascertain the truth of $P_k(k)$. This is essentially the argument put forward by Lucas (1961), but it is not yet the desired contradiction, since the mathematician can have no means of knowing what F *is*, let alone be convinced of its validity as a means of ascertaining truth. We shall need a broader argument than this.

Suppose that the ways that human mathematicians form their conscious judgements of mathematical truth are indeed algorithmic. We shall try to reduce this to an absurdity. Consider, first, the possibility that different mathematicians might use *inequivalent* algorithms to decide truth. But it is one of the most striking features of mathematics (perhaps almost alone among the disciplines) that the truth of propositions can actually be settled by abstract argument. A mathematical argument that convinces one mathematician – provided that it contains no error – will also convince the other, as soon as the argument has been fully grasped. This also applies to the Gödel-type propositions. If the first mathematician is prepared to accept all the axioms and rules of procedure of a particular formal system as giving only true propositions, then he must also be prepared to accept its Gödel proposition as describing a true proposition. It would be exactly the same for the second mathematician. The point is that the arguments establishing mathematical truth are *communicable*.

Thus we are not talking about various obscure algorithms that might happen to be running around in different particular mathematicians' heads. We are talking about *one* universally used (putative) formal system that is equivalent to all the different mathematicians' algorithms for judging mathematical truth. Now this putative "universal" system, or algorithm, cannot ever be known as the one that we mathematicians use to decide truth. For if it were, then we would construct its Gödel proposition and know that to be a mathematical truth also. Thus, we are driven to the conclusion that the algorithm that mathematicians actually use to decide mathematical truth is so complicated or obscure that its very validity can never be known to us.

But this flies in the face of what mathematics is all about! The whole point of our mathematical heritage and training is that we do not bow down to the authority of some obscure rules that we can never hope to understand. We must *see* – or, at least in principle see – that each step in an argument can be reduced to something simple and obvious. Mathematical truth is not a horrendously complicated dogma whose validity is beyond our comprehension. It is something built up from simple and obvious ingredients – and when we understand them, their truth is clear and agreed by all.

To my thinking, this is as blatant a *reductio ad absurdum* as we can hope to achieve, short of an actual mathematical proof. The message should be clear: Mathematical truth is not something we ascertain merely by the use of an algorithm. I believe, also, that our *consciousness* is a crucial ingredient in our comprehension of mathematical truth. We must "see" the truth of a mathematical argument to be convinced of its validity. This "seeing" is the very essence of consciousness. When we convince ourselves of the validity of Gödel's theorem we not only "see" it, but by so doing we reveal the very nonalgorithmic nature of the "seeing" process itself.

The strong-AI view, on the other hand, envisages that human judgements of (mathematical and other) truth must arise out of some kind of *natural selection of algorithms*. Even in the absence of arguments in favour of some nonalgorithmic ingredient in conscious thinking, such as the one given above, there are serious difficulties with the picture whereby algorithms are supposed to improve themselves in this way. It would certainly not work for normal Turing machine specifications, since a "mutation" would almost certainly render the machine totally useless instead of altering it only slightly. Something much more robust would be needed – such as the actual *ideas* underlying the specification! Moreover, a selection process that relies only on the output of an algorithm (such as natural selection implies) would be hopelessly inefficient.

What do I believe is actually taking place with our conscious perceptions? I am venturing a suggestion (partly as a result of my own experiences with the communication of mathematical ideas) that consciousness represents some kind of contact with the timeless Platonic world of mathematical concepts. Proper communication between mathematicians can take place only when each individually makes this contact. Moreover, the "nonalgorithmic part" of Plato's world is, by comparison with the part where algorithms and computation reside, easily the most subtle and fascinating part. But mathematical insight is only one area where the role of consciousness is important, and it is chosen in my descriptions simply because it is here that the nonalgorithmic element can be made most precise. *Any* conscious thinking, even that which I believe must occur in (at least some) animals would, at root, have to be the same phenomenon.

The relation between the real physical world and the Platonic one is a mysterious one. The very existence of the SUPERB physical theories begins to give the seemingly "solid" matter of our experiences an apparently nebulous "mere mathematical" Platonic existence, whereas examples such as the Mandelbrot set seem to make the Platonic world more concrete. Might the two worlds be, in some sense, actually the *same?* The whole issue of consciousness seems to me to require a much deeper investigation of the relation between the physical world and Plato's world than has been achieved so far. Even the standpoint of strong AI requires the Platonic existence of algorithms as the "home" for our conscious feelings, since the algorithms' physical embodiments are supposed to be irrelevant. Viewed in this way, my own standpoint does not differ so much from that of strong AI, except that I believe that the concept of an algorithm is far too limiting to accommodate the immensely subtle and fundamentally important phenomenon of consciousness.

The possible identification, in some sense, of the physical world with the Platonic one raises many issues. There is the possibility of *strong determinism* (according to which the actual history of the universe might be mathematically fixed in its entirety). There are also the various versions of the *anthropic principle*, which I discuss briefly in *Emperor*. In its (more dubious?) *strong* form there could even be *alternative universes*, the one in which we find ourselves being the one that happens to allow consciousness.

Suppose we accept that there may indeed be a nonalgorithmic element in the physics that spans the gap between the quantum and classical levels. How might the conscious brain be making use of this? Here the speculations become more tentative. A plausible place to look is in the phenomenon of *brain plasticity*. Connections between neurons sometimes occur at *dendritic spines*, which can apparently grow or shrink, and significantly affect connection strengths, in seconds or less. This growth or shrinkage should, strictly speaking, be considered as taking place at the quantum level: No single one of the vast array of possibilities has actually occurred; *quantum superpositions* of many different combinations of connection strengths must be simultaneously involved. Thus, the brain embarks not on just one "calculation," but on many simultaneously. Only when the large-scale effects of these simultaneous calculations differ by something that reaches the "one-graviton level" – or whatever CQG tells us is appropriate – would one or another "conscious thought" emerge. Moreover, this "emergence" would have to involve some nonalgorithmic ingredient.

I suggest a possible analogy with *quasicrystal* growth. These puzzling objects resemble crystals, but possess crystallographically forbidden symmetries, usually *fivefold*. Fivefold ("almost") symmetry can occur with certain tiling patterns, however, and it seems likely that the arrangement of the tiles in such patterns will give the essential clue as to the arrangement of atoms in a quasicrystal. If so, there would have to be an essentially nonlocal aspect to the way the atoms are arranged. I am of the opinion that this cannot be properly achieved by adding atoms to the assembly one at a time. Quantum superposition of many different simultaneous arrangements must be involved; only when the classical level is reached does a particular one of these arrangements get resolved out. (It should be recalled that the general tiling problem is a nonalgorithmic one, although the particular tilings that are likely to be relevant for such quasicrystals are actually algorithmic – though nonlocal.)

Finally, I refer in *Emperor* to certain experiments (by Kornhuber and by Libet) that indicate some very puzzling aspects of the actual time at which conscious feelings (active and passive, respectively) seem to take place. [See Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4)1985.] There is also something very puzzling about the fact that consciousness is the one known phenomenon for which time needs to *flow* at all. In the descriptions of modern physics – except, perhaps, for the (disputed) behaviour of **R** – all we have is a "static" space-time. I believe that these are strong indications that we need a radical revision of our present picture of space-time. For such a revision, we would certainly need something like the putative CQG.

Despite all this technicality (and also physical speculation), there remains the fact that in some sense it is "obvious" that mere computation cannot evoke consciousness – an obviousness that a child can see. Yet our science has driven us to accept that we are all but small parts of a universe governed by precise mathematical laws. In *The Emperor's New Mind*, I have attempted to show that there is a way out of the dilemma: Mathematical precision does *not* imply computability.

NOTE

1. I have been recently informed by Leonore Blum that this conjecture is actually true, if one uses the definition of noncomputability for complex numbers due to Blum et al. 1989.

196

AUTHOR'S RESPONSE – THE NONALGORITHMIC MIND

Behavioral and Brain Sciences **13** (1990), no. 4, 692–705

Without the actual individual 37 critical reviews (Open Peer Commentary), these responses of mine inevitably give a somewhat one-sided perspective. At least they spell out some aspects of my own viewpoint more completely than earlier. For the 37 commentaries themselves, the reader is referred to the journal (as given in the running head above, on pp. [655]–[691]). There were five further commentaries, which were basically responses to the present remarks, and my own responses to those constitute Chapter 225.

Author's response – The nonalgorithmic mind. *Behavioral and Brain Sciences* **13** (1990), no. 4, 692–705

Author's Response

The nonalgorithmic mind

Roger Penrose
University of Oxford, Mathematical Institute, 24-29 St. Giles, Oxford, England OX1 3LB

I think it is fair to say that the least complimentary of these thirty-seven commentaries are also the ones containing the most serious misunderstandings of what I was trying to say in *The Emperor's New Mind* (henceforth *Emperor*) They add up to at least one very sobering – and evidently valid – censure: The book was not nearly clear enough!

It is perhaps not surprising that the most negative remarks come from the AI community. Let me begin by saying that I intended no criticism whatsoever of what they are doing or of their *striving* to simulate intelligence or other activities of human or animal nervous systems. I had hoped this would be clear from a detailed reading my book. Although the title might be interpreted as suggesting that the proponents of AI are (like the fraudulent weavers in Hans Anderson's "The Emperor's New Clothes") trying to "put one over," on the general public, certainly no such insinuation was intended by me. (**Glymour & Kelly** say that my title is "almost the only rude thing about the book"!) I was aware, when I chose my title, that such an interpretation was possible, but I had not expected that some AI proponents would be sensitive enough on this point to take such an interpretation seriously. I hope that they will accept both my dissent from such an interpretation and also my unreserved apology, if they believe that some might be misled into thinking that I viewed AI proponents as being in any way dishonest in their intentions – which I emphatically do not.

Although some criticisms came from what I shall call the "soft" side (people who believe that mental phenomena cannot be accounted for in terms of physical explanations: possibly **Niall**, **Roeper**, **Taylor**, **Varela**, and **Zytkow**), it was clear that most would come from the "hard" side (those with at least some admitted sympathy for the position of "strong AI": **Boyle**, **Breuel**, **Dennett**, **Doyle**, **Eagleson**, **Glymour & Kelly**, **Johnson**, **Ettinger & Hubbard**, **MacLennan**, **Madsen**, **McDermott**, **Perlis**, **Smithers**, **Tsotsos**, **Waltz & Pustejovsky**, **Wilensky**). I in fact believe that there are strong arguments on both sides, as I tried to explain in my book. Those on the "soft" side are right in their belief that there is no evidence whatever to suggest that "feelings," or "qualia," can be described in physical terms at all, let alone be evoked by the mere carrying out of a calculation; while those on the "hard" side are equally right to point out how intimately our feelings and perceptions are associated with the states of particular physical objects, namely brains, while all physical objects are, as far as we know, like patterns of "information" (?) governed in full detail by marvellously precise mathematical laws. The very strengths of both arguments had seemed to present us with a paradox. The way out of this paradox, in my opinion, must lie in our obtaining much deeper insights into what physical laws actually *are* and in unearthing those aspects of physics that must evade computational descriptions and begin to allow room for experiential phenomena. It is towards such goals that I have attempted, with my book, to guide our future thinking. Let me now make some comments addressing the arguments on both sides.

Why physics?

Those on the "soft" side might wonder why I even bothered to write the book at all, since it is surely "obvious" that subjective experiences, or "qualia," are not things that can ever be reduced to physics (i.e., to "mass, length, and time," in **Niall**'s words; see also **Zytkow**'s account). So what light can possibly be shed on this question by bringing in physics at all, let alone CQG ("correct quantum gravity")? To this I must reply, "precious little, if anything, so far" as they might expect. I am certainly not claiming to have a theory of the mind – and I am rather surprised that some of my critics (e.g., **Boyle**) seem to be chastising me for not presenting one! It would be presumptuous in the extreme for me to claim to have solved a fundamental puzzle that had remained unsolved for some three thousand years despite its having been the subject of deliberations of some of the deepest thinkers since antiquity. Introspection can certainly yield significant information but introspection has not really got us very far with regard to the basic question of what perception *is*, or of why we perceive at all.

It seems to me that we shall never be able to make progress toward understanding these issues until we have a much better picture of what physical reality is actually all about (and with reference to **Glymour & Kelly**'s remarks, I do not see how we can do without physical realism, whatever view one might hold about the relation between the mind and computation). It is in science, and above all physics (profoundly supported by mathematics), where enormous strides in our understanding of this world – the universe in which we (and our minds) actually appear to live – have been achieved. It is in physics that we have at last found SUPERB theories, and although these theories are still very incomplete and do not yet shed any significant light on what minds actually are, it is through the back doors of physics and mathematics, in my opinion, that our assault on the problem of mind must ultimately be mounted. The SUPERB theories which have emerged in this century, namely, relativity and quantum theory, have actually moved us some good way from the rather sterile-seeming "mass, length, and time" characterization of classical physics to something much more subtle. I have no doubt whatever that there are more profoundly subtle theories (still to be found) underlying both relativity and quantum theory, and that the more we understand about the nature of our world, the less sterile the physics of this world will seem – and eventually the actual role that consciousness and perceptions play in our universe will begin to be discerned. Despite what some of my critics (**Stanovich, Waltz & Pustejovsky**) seem to be implying, I certainly do not believe that CQG (correct quantum gravity), when it finally comes to light, will in itself solve the problem(s) of mind. But I have tried to present arguments in my book in support of my view that some (nonalgorithmic) physical theory of this kind (very probably CQG) is a necessary prerequisite if a scientific theory of mind is ever to be achieved.

Simply finding a (necessarily algorithmic) "physical grounding for information processing" (**Boyle**) in terms of existing physics will not be enough. Nor do we know, despite what **MacLennan** seems to imply, that physical behaviour is necessarily algorithmic just because described by differential equations. (**Breuel** even seems to take the opposite view, see later!) A good part of my purpose was to stress that computability in physics is not at all the same thing as determinism (**Eagleson** and **Taylor** please note).

Mathematical insight and Gödel's theorem

Let us now turn to the question of why I think that an *algorithmic* theory like strong AI can never provide a theory of mind. A large part of my argument is based on Gödel's theorem, but a good many of those who criticized me from the "hard" side (**Dennett, Doyle, Glymour & Kelly, MacLennan, McDermott, Perlis, Tsotsos, Wilensky**), and also several others (**Boolos, Butterfield, Chalmers, Davis, Hodgkin & Houston, Kentridge, Manaster-Ramer, Savitch & Zadrozny, Mortensen, Roskies**), have objected, and even claim to have located errors in my argument ("quite fallacious," Chalmers; "wrong," Mortensen; "lethal flaw" and "inexplicable mistake," Manaster-Ramer, Savitch & Zadrozny; "old and fallacious argument," Perlis; "deeply flawed," Davis; "invalid," Kentridge) – while only a few seem partially to support my type of argument (perhaps **Gilden & Lappin, Higginbotham, Lutz, Niall**) on this issue. As I shall explain in a moment, there is actually nothing wrong with my argument – and I make no apology for the fact that much of it is "old," certainly appreciated by Gödel himself already in the 1930s and never properly refuted since (cf. Gödel 1986; 1990; and as cited in Rucker 1982, p. 171; also Nagel & Newman 1958) – except that I evidently did not present my case forcefully enough, nor did I anticipate the most likely misunderstandings. All my adverse critics on this topic have jumped to conclusions and, in one way or another, have missed the point of what I am trying to say. None seem to have grasped the full import of the Gödelian argument. The fault is mine: I should have explained things more clearly.

Most of these critics have pointed out that the Gödel proposition ($P_k(k)$ (which I shall subsequently denote simply by $G(F)$ here), constructed from some formal system **F**, cannot be "seen" to be true unless the system is already known to be consistent. (Actually the particular Gödel-type proposition I construct does not depend on the consistency of F, as my critics maintain, but, as the argument is given, on what is called its "ω-consistency." This is stronger than actual consistency, asserting that if $\sim\forall x[Q(x)]$ is a theorem in F, then it must not be the case that all the propositions $Q(0)$, $Q(1)$, $Q(2)$, $Q(3)$. . . . are theorems. This depends upon the symbol " ~ " and, more particularly, "∀" actually meaning what they are supposed to mean: "not" and "for all" respectively! The distinction between consistency and ω-consistency is not important for us here – and in fact one can get away with assuming much less for F – so let it pass for the moment.) My point is not that, in any particular case, $G(F)$ is true or "seen" to be true (or false or "seen" to be false). My point is that the *deduction* of $G(F)$ from F is (and is "seen" to be) a perfectly valid procedure. Thus, if we are already accepting F as a sound system, then we must accept $G(F)$ as being true also. (The word "sound" is used by logicians for systems whose theorems are actually true in the intended interpretation of the symbols.) If we are prepared to use the procedures of F as sound methods of inference then we must accept the passing of F to $G(F)$ as another sound method of inference. This is not unlike any other form of reasoning in mathematics. One does not directly perceive a "new" mathematical truth in isolation. What one does is to see that the new proposition follows, in some way, from the body of mathematical understanding that has already been established. It is in this sense that one "sees" the truth of $G(F)$, just as one does in constructing any mathematical argument for obtaining new truths from old. But the Gödel method lies outside the procedures that have been actually laid down within F itself. I am not concerned with how we may (or may not) already "know" that the procedures of F are sound – that is part of what is assumed – but with the remarkable fact that if we are accepting F as sound, then we must also accept $G(F)$ as true.

Some of my critics (**Boolos, Chalmers, Davis, Perlis**) have drawn attention to some particular formal system (e.g., the standard Zermelo-Fraenkel system **ZF** – or a modification thereof, such as **ZFC** or **ZFM** – cf. Boolos, Davis), claiming that we do not know whether or not it is consistent. Thus we cannot be sure that its Gödel proposition is actually true. (Inconsistency is just one particular type of manifestation of a system F not being sound. There are many others, even apart from the ω-inconsistency actually relevant here. The system might be consistent, or even ω-consistent, whilst some of its implications might be false; but my critics seem to have paid attention only to the possibility of inconsistency.) If F is inconsistent, then any proposition that can be described within the system is derivable using the rules of F, irrespective of whether it is true or false. Thus, in particular, G(F) is itself a theorem! In fact, the Gödel proposition for any system F is always either true as a statement in mathematics or it is a theorem of F. Of course, the latter are the situations in which F is useless as a means of deriving mathematical truth. (If F is inconsistent, then, in particular, 2 = 3 is a theorem also.) But I simply do not understand how mathematicians of a formalist disposition[1] (Boolos, Davis) can claim that mathematical argument within some particular formal system F are more acceptable than the mathematical argument that derives G(F) from F.

In his criticism, **Boolos** asserts that the systems generally used in mathematics (e.g., ZF) are good approximations to mathematics. (Approximations to what?! Boolos is revealing himself to be a good Platonist after all, since he evidently believes that there is something there, other than the formal systems themselves, that the formal systems are supposed to be approximations to.) But if ZF turns out to be inconsistent, then it is no approximation to mathematics at all since 2 = 3 would then be a theorem according to the rules of ZF (**Mortensen** take note!) – though the *rules* of ZF might approximate the rules of some consistent formal system which does approximate actual mathematics.

To reiterate my central point: I am not asserting that, in any particular case of a formal system F, we need necessarily be able to "see" that G(F) is true; but I am asserting that the "Gödelian insight" that enables one to pass from F to G(F) is just as good as a mathematical procedure for deriving new truths from old as are any other procedures in mathematics. This insight is not contained within the rules of F itself, however. Thus, F does not encapsulate all the insights available to mathematicians. Such a limitation applies to any formal system whatsoever. I nowhere claim that the mathematicians' insights would enable them, in principle, to resolve *any* mathematical question (**Dennett, Doyle, Roskies**). I merely claim that the insights which *are* available to mathematicians are not formalizable.

MacLennan (and also **Wilensky**) refers to a distinction between mathematics and metamathematics, the deduction of G(F) from F being, he asserts, metamathematics, not mathematics. But this distinction only really exists if one adheres to a purely formalist view of mathematics. To a Platonist (or a "common-sense mathematician"), metamathematics – which is reasoning *about* formal systems (rather than within a formal system) – is still mathematics, on a par with any other form of mathematical reasoning.

I should make a remark about formal systems, such as ZF, ZFC, or ZFM, which refer to reasoning about infinite sets. These sets are allowed to be rather enormous; one might indeed have reasonable doubts about their existence or about the system's consistency (at least in the case of ZFM). One might even encounter differences of opinion amongst mathematicians as to whether or not to accept reasoning which involves such infinite sets. But one can avoid much of the argument about such possibly controversial issues by concentrating on just the small area of mathematics consisting of propositions in number theory. In fact, it is sufficient merely to consider propositions of the special form

[Q]: "Q(n) is true for all natural numbers n"

where Q(n) is some given computable (recursive) true/false arithmetical property of the natural number n (i.e., there is an algorithm for deciding, for any particular n, whether Q(n) is actually true or false – so Q(n) is an always terminating Turing machine action).[2] The point about this is that the Gödel proposition G(F), as presented in *Emperor*, is explicitly of this particular form, so we do not have to consider anything set-theoretically more obscure than statements of this kind.

How do mathematicians decide the truth or falsity of such propositions? They might decide to use the rules of some formal system F that (like ZF) refers to mathematical structures (say, enormously infinite sets) that go far outside the concepts of ordinary arithmetic, and in doing so, might introduce some doubt (e.g., see **Davis**) as to their consistency, let alone their soundness. Nevertheless, for any given system F, there will be propositions of the form [Q] that can be "proved" using F and others that cannot. If F is inconsistent, then the system F would be, strictly speaking, useless. The decision as to whether or not to trust a formal system F as being sound is normally the job of those logicians concerned with the details of that system. But if I choose, myself, to use the rules of F in some calculational way – say, to decide the truth of some proposition of the form [Q] – then, to know whether I am to trust a result derived from these rules, I, also, should persuade myself that F is *sound*. This is a much stronger requirement (i.e., actually *using* the rules of F) than I would need if I were to use my Gödelian insight to deduce G(F), since for that I need only know that F is ω-consistent. All mathematical knowledge requires mathematical insight and understanding. This applies to the judgements I might apply in trying to decide whether or not F is sound; it also applies to the Gödelian insight that enables me to deduce G(F) from F. The Gödelian insight is just one kind of mathematical insight amongst many. It is emphatically not something given only to a "chosen few" such as myself (as **Glymour & Kelly** seem to claim with some irony). Indeed, it is an insight simple enough even for me to give a reasonably adequate description of it in my (semipopular) book! It is something noncontroversial and, in principle, accessible to all of us, unlike (apparently) the insights lying behind the validity of certain particular systems like ZF or, more especially, ZFM. The important thing to realize about the Gödelian insight is that it derives from an (at least partial) understanding of the *meanings* of the axioms and rules of procedure of F, and also of the actual meaning of G(F). (Most particularly, it derives from an understanding that the symbol "∀" actually means "for all natural numbers," since this the essence of ω-consistency. Gödel's theorem in the form that I have presented it illustrates the fact that the actual meaning of "∀" cannot be fully encapsulated by formal rules. **Mortensen** is puzzled how a "finite" mathematician can comprehend such things as hyperreal numbers. But the puzzle is more simply based than this, since it applies already to the natural numbers.) One cannot regard F as providing the rules merely of some "game" for manipulating meaningless symbols (as the strict formalist might lead one to believe). Thus Gödel's theorem has another profound implication (**Garnham** take note; also, I think **Roskies** grossly underestimates this point, whereas **Smithers** does not even seem to recognize it!): *Semantics cannot be formalized.*

Author's response – The nonalgorithmic mind. *Behavioral and Brain Sciences* **13** (1990), no. 4, 692–705

Are mathematical insights algorithmic?

It is important to realize that "formalizable" and "algorithmic" are virtually synonymous terms. (The "proving of theorems" – say of statements of the form [Q] – within some formal system **F**, is certainly something algorithmic whereas, conversely, any algorithm that generates "theorems" can be adjoined appropriately to the axiom system of predicate calculus: a standard set of rules of logical inference). Thus we can directly apply the Gödelian insight to any algorithm that purports to be generating (say [Q]-type) mathematical truths. (Note: apart from considerations of complexity theory, a truth-generating algorithm is formally equivalent to a truth-proving algorithm – despite what **McDermott** says!) If the algorithm in fact generates only mathematical truths then we can obtain a new mathematical truth by using the Gödelian insight – a truth that is not obtainable by the given algorithm (together with predicate calculus). The original algorithm is thus limited in what it can achieve: It is unable to incorporate the particular insight

$$\text{"}\mathbf{F}\text{ sound"} \Rightarrow \text{"}G(\mathbf{F})\text{ true"}$$

that Gödel has revealed to us. Moreover, as soon as we have seen that G(**F**) is true we realize that we have not finished with the insight, because "G(**F**) true" can now be adjoined to "**F** sound" (etc. etc.).

Some of my critics (**Davis, Doyle, Kentridge, MacLennan**) have objected, claiming that the Gödelization procedure could itself be automated. But this is not so if by "procedure" we mean the above "insight." If the Gödelization procedure as a whole *could* be automated, then we could apply this procedure to the very system that the automation would yield, and hence obtain a new Gödel proposition that lies outside the scope of that automation. In detail, the way that this can come about is quite subtle, and it is perhaps not surprising that people often feel that they could automate Gödelization. The Gödelian insight is a very slippery character, who can flee in a different guise as soon we feel that we have got him algorithmically cornered.

Let us see how this can come about. Suppose that we have some formal system **F**. We wish to adjoin to **F** another rule which might seem to encapsulate the Gödelian insight: "If {A, B . . . D} is any finite[3] set of theorems, then the Gödel proposition obtained from them as axioms (and the rules of procedure of **F**) is to be also a theorem." This gives a new system $\mathbf{F}^*$ that is much broader than that obtained by simply adjoining G(**F**) to the axioms of **F**, since it allows Gödelization to be applied again and again, as many times as we might wish, always within this one system. If **F** is sound, then so also will be $\mathbf{F}^*$. But $\mathbf{F}^*$ still has not really incorporated the Gödelian insight because the proposition $G(\mathbf{F}^*)$ will now lie outside the scope of $\mathbf{F}^*$, and $G(\mathbf{F}^*)$ will be true if **F** is sound (and will be seen to be true so long as **F** is seen to be sound). Of course we might now perceive how to automate the procedure $\mathbf{F} \Rightarrow G(\mathbf{F}^*)$, rather than just $\mathbf{F} \Rightarrow G(\mathbf{F})$, but again, that very perception would allow a new Gödelization procedure that lies outside the scope of that new automation.

The slipperiness of the nonalgorithmic nature of the Gödelian insight lies partly in the fact that we are never quite sure what it is that has slipped through the algorithmic net, though something always does! One might argue that what has slipped through it is the very perception of how one actually formulates something in an algorithmic way, in the knowledge that the suggested algorithmic procedure actually yields a truth (or a set of truths). In particular, the algorithmic procedures that **Davis** refers to whereby the consistency of a set of axioms can be coded into the nonsolubility in integers of some polynomial equation $P = 0$ involve the knowledge and insights (very ingenious ones, at that, for which Davis himself was partly responsible!) that the insolubility of that equation actually does encode the consistency statement it is supposed to encode. There would also be numerous *unsound* algorithmic procedures of this type, but we need (nonalgorithmic) insights to tell us which procedures are in fact sound!

It seems to me that this is the same kind of slipperiness of the roles that conscious understanding and insight play in any kind of mental activity – especially as they relate to the procedures whereby one might try to program a robot to behave in certain humanlike ways. As soon as we have seen how to turn the implications of a certain insight into a set of rules, we can make a robot act according to those rules. There seems to be nothing barring our being able to automate anything we can precisely formulate by such rules – and when it is automated, the robot can carry out the implications of that particular automation far faster and more accurately than we can ourselves. Yet the very insight itself has not been automated. Nor has the insight that we have used a particular insight to program the robot. Insights in general cannot themselves be automated – as is illustrated by the fact that the full Gödelian insight, in particular, can never be automated. The understanding of what that insight actually is doing will allow us to do better, in some ways, than the robot that we have constructed by using it.

Learning to perform some human action that initially requires some conscious understanding appears to be a very similar phenomenon. Once learnt, that action can be relegated to unconscious (perhaps cerebellar rather than cerebral) control. This unconscious control seems much more like that of a programmed robot – and as with an electronic computer's action, it can be carried out much faster and more accurately than can action under conscious control – while unconscious control does not appear to embody the actual understanding the initial consciousness was needed for. What is the actual role that consciousness plays? It seems to be a subtle – and even a slippery role – like that of the Gödelian insight. This slipperiness is in the nature of all nonalgorithmic understanding.

I should emphasize that there is nothing "paradoxical" in the Gödelian insight, unlike (as **MacLennan** seems to be suggesting) the case of the "liar" paradox. People often seem to think that the self-referential aspect of Gödel's theorem is its essential content. This is extremely misleading. The relation of Gödel's theorem to the liar paradox is largely a historical and motivational one. Most nonalgorithmic classes of problem have no evidently self-referential aspect at all. Of course human language is imprecise, and it does allow inconsistent selfreferential statements like, "This statement is a lie," to be made. (Compare **Wilensky**'s self-destructing assertion that my deductions from Gödel's theorem and their relevance to

AI are "uncontroversially wrong"!) This imprecision is just as well, for without it, we should be constrained to converse within some specific formal system, and discussion of matters such as the Gödelian insight itself (and other reflection principles) – let alone semantics in general – would become impossible.

Kentridge refers to my version of the Gödelization procedure (as given in *Emperor*) according to which one can, by the exercise of understanding, "outdo" any given algorithm H for testing the stopping of Turing machine actions.[4] He says that with this argument (as with the Gödel argument), all I have shown is that humans can do better than one particular algorithm, not better than all algorithms. Not so! To say this is to miss the point of what one is achieving with a *reductio ad absurdum*, of the kind I have been presenting. (**Chalmers** and **Wilensky** also take note – and despite **Zytkow**'s friendly warning, I do not lose the "outdoing an algorithm" game!) My argument is of the form: "Given any particular algorithm, that algorithm cannot be *the* procedure whereby human mathematicians ascertain mathematical truth. Hence humans are not using algorithms at all to ascertain truth." Recall, for comparison, Euclid's argument that there is no greatest prime number. This is also by reductio ad absurdum: "Given any number n that purports to be the greatest prime, n! + 1 has a prime factor larger than n, so n was not the largest prime after all. Hence there is no greatest prime." As with the algorithms above, **Kentridge** would appear to be claiming that all Euclid has shown is that for any particular number n, one can find a larger prime. But this certainly *does* suffice to show that there is no largest prime! The argument with the algorithms is similar. Given any algorithm that purports to be the one that humans knowingly use to ascertain mathematical truth, one can find a more powerful algorithm that we (humans) can see will also ascertain mathematical truth. This likewise suffices to show that humans do not ascertain mathematical truth by means of any knowable algorithm.

A more serious issue is that humans might be using an algorithm whose validity is in fact unknowable to them, and this is indeed the very point I was trying to address in the final chapter ("disappointing" to **Tsotsos**) of *Emperor*. Only a very few of my critics appear even to have noticed that my argument is not the same as that of Lucas (1961). Rather than addressing, with Lucas, the possibility that some particular individual might be acting according to an algorithm, I refer instead to a putative algorithm that the mathematical community *as a whole* is supposed to be using. It is rather more believable that any particular individual might be using some horrendously complicated unknowable algorithm than that the mathematical community as a whole is using one. The existence of an algorithm X, according to which mathematicians as a whole are supposed to act, would place an absolute limit on what human mathematicians are able to achieve. The algorithm (or formal system) X, or at least the soundness (or merely ω-consistency) of X, would have to be something unknowable to human mathematicians. Otherwise G(X) would be seen to be a mathematical truth, although it is inaccessible to X: a contradiction.

I never claimed that the existence of X is a mathematical impossibility, but it would seem to be exceedingly unlikely. Appealing to a "horrendously complicated unknowable algorithm" is totally at odds with the actual way mathematics is done. As I said in *Emperor*, mathematics is "built up from simple and obvious ingredients" although sometimes, as when one uses a Gödel type insight for a reflection principle[5] these ingredients are sometimes quite subtle – though still "obvious"! – and their very "obviousness" may require some genuine appreciation of the underlying meaning of the symbolism used. (It should have been clear that I am not claiming, as **Davis** and perhaps **Chalmers** seem to be suggesting, that mathematical arguments are themselves always "obvious." That would be very far from the truth, since they are sometimes exceedingly complicated. It is just that such arguments are, at least in principle, built up from such "obvious" ingredients. Moreover, I don't even need any outrageously subtle "obvious" ingredients. The Gödelian insight alone will do (**Chalmers** and **Higginbotham** take note.)

Of course, it might be the case (and I suppose that the following is the sort of thing that the AI people have in mind, cf. **Hodgkin & Houston**, also **Butterfield**) that mathematicians have the *illusion* that all the time they are appealing to such "simple and obvious ingredients" whereas in fact they are actually using the horrendous X, where X has somehow arisen by natural selection. This seems to be an implication of the sort of picture being put forward by **Dennett** in particular, although I am not altogether sure that I have understood his point completely. Somehow, X is supposed to have arisen because our remote ancestors were clever at designing mammoth traps and the like, and now some of their descendants have found that X is also good for discerning obscure mathematical truths! I find such a picture quite implausible. It is creatures who can *understand*, and who can comprehend *meaning*, be it mathematical or be it conceptually valuable for the construction of mammoth traps, that natural selection has favoured. Understanding and meaning are nonalgorithmic (as the Gödelian insights demonstrate) and they are mental attributes that clearly seem (to me, at least) to be particular manifestations of conscious contemplation.

Perlis makes the very curious suggestion that Fermat might have worked according to the system ZFC (as his "X") and that that was how he discovered mathematical induction. But that would imply that the great Fermat was intellectually incapable of understanding the axioms of ZFC – and certainly incapable of believing them!

I cannot agree with **Butterfield** that consciousness is just "baggage" from the point of view of natural selection, like the weight of the polar bear's coat. I argue that consciousness is a necessary ingredient of understanding and insight, so I believe that it must, in itself, have been a very positive factor with regard to natural selection. I should emphasize that I strongly believe in natural selection, although some (e.g., **Hodgkin & Houston**) seem to have taken a certain passage in *Emperor* (p. 416), to the effect that natural selection appears to work better than it "ought" to, as implying that I have significant doubts. I was referring only to the remarkable way in which the actual physical laws, with their propensity towards forming complex and highly organized structures (like quasicrystals or Frank-Casper phases of crystals) seem to fit in with natural selection extraordinarily well, and making the whole process work better than it ever could if the laws themselves were not so appropriate. **Breuel** (as well as **Hodgkin & Houston**) refer to "fault tolerant" algorithms that are more robust than ordinary Turing machine specifications. I agree that such things would be preferable from the point of view of natural selection, but I still do not see how the horrendous X could possibly be selected for in this kind of way.

Dennett refers to the algorithms that chessplaying computers use, these being "extremely good" rather than "correct" algorithms for playing chess. I think that he (and, in effect, **Taylor** also) is suggesting that mathematicians are likewise using such an "extremely good" rather than "perfect" algorithm to decide mathematical truth. But I really do not see how Dennett gets around the above argument. His proposal does not at all explain the types of insight mathematicians are using all the time, which are *correct* but nonalgorithmic rather than *approximate* but algorithmic. Of course, as **Boolos, Chalmers, Davis, Garnham, Hodgkin & Houston, Lutz, Perlis,** and **Roskies** point out, mathematicians sometimes do make mistakes (although Cantor was *not* shown by Russell to be wrong), but the mathematical community as a whole makes extraordinarily few; and this would be totally inexplicable from the "naturally selected X-algorithm point of view" given the extraordinary abstractions away from actual experience involved in most of pure mathematics. Any approximate algorithm for generating all the mathematical truths we know would have to be incredibly complicated (and in no way related to the kind of thing creatures need to survive), whereas mathematical insights are, at root exceedingly simple.

It seems that Turing himself thought that somehow it was the fact that humans make mistakes that got around the seeming paradox of an algorithmic device (as was his view of the brain) being able to appreciate the truth of the nonalgorithmic Gödelian insight (see Hodges 1983). To me his proposed way out seems exceedingly unlikely, for essentially the reasons outlined above; and it is hard to see that introducing inaccuracy is what makes the brain work better! I have said that the imprecision of human language allows it greater scope than otherwise, but language only works at all because of the underlying power of the mind that allows precise sense to be extracted from imprecise language (e.g., from an informal description of the Gödel procedure: **Garnham** take note). Gödel (cf. 1990, p. 297) found himself forced into a different extreme, namely, that the "mind" was only loosely to be associated with the brain, and was not limited by the brain's finiteness. The fact that these two great thinkers found themselves driven to take such improbable-sounding positions is an indication of the genuine seriousness of the problem which must be faced. Neither Turing nor Gödel seems to have considered the possibility of nonalgorithmic physical action operative in the brain. If this also sounds improbable to some on the face of it, they should at least appreciate why something remarkable is actually needed!

I should make it clear that in taking the argument outward from single mathematicians to the mathematical community as a whole, I am not allowing that a nonalgorithmic entity (the "mathematical community") might somehow arise from a collection of algorithmic ones (as some indeed suggest individual mathematicians might be). On the other hand, at one stage down from this, **Chalmers** (and perhaps **Doyle** and **Perlis**) seem to be suggesting that humans might manifest nonalgorithmic behaviour on a large scale while being constructed from entirely algorithmic (microscopic) ingredients. In this view, it would be possible in principle to simulate the behaviour of the ingredients using a complicated-enough computer program and make it look as though the simulation was behaving nonalgorithmically on the large scale. This is basically a version of the same suggestion as Dennett's above, where a "horrendous X" algorithm is supposed to underlie all human insights and understanding. My (main) arguments against this are just as before. Moreover, despite what Perlis seems to be claiming, it seems exceedingly unlikely (though perhaps not totally impossible) that anything usefully nonalgorithmic can ever arise out of algorithmic basic ingredients in this way. For example, "chaos" (referred to by **Mortensen,** though I am not altogether clear about what his point is, and also by **Waltz & Pustejovsky**) does not seem to be "usefully nonalgorithmic," as I discuss at length in ***Emperor***, but perhaps this is a question that is worthy of some more serious study.

Are there any limitations, in principle, to what human mathematicians can achieve? I would not know the answer to this question, but (despite what **Dennett, Doyle,** and **Roskies** appear to be implying) I do ***not*** assume that there are no such limits – although it does seem to me to be at least possible that in principle there are no such limits. There are obviously limits in practice, since, for example, no human will ever be able to multiply together each pair of numbers below $10^{10^{1,000,000}}$. (Even a computer could not do this, but it would, of course, be much better at this sort of thing than a human!) Limits of this kind are not what is at issue, for the human mathematician knows how to perform the multiplications in principle, and this is all that is required. To see that human thinking is nonalgorithmic, we need only refer to what can be achieved in principle. Most mathematical thinking is of the "in principle" kind in any case.

What do I mean by "algorithm"?

I am puzzled that various of my critics seem to be uncertain as to what I (or others) actually mean by the term "algorithm." **Eagleson, MacLennan, Smithers, Taylor, Tsotsos,** and **Wilensky** seem to be implying that I use the term in more than one sense, or in a sense in which I do not mean it. By an algorithm I always mean something equivalent to the action of a Turing machine. I had thought that this was standard terminology, and it is certainly what I adhere to in my book. Any modern general-purpose computer is, in effect, a (universal) Turing machine. There is of course the fact that, strictly, a Turing machine's tape is infinite – or, at least potentially unlimited – but this (acceptable) idealization does not seem to be what is at issue. Any algorithmic action can, with this proviso, be carried out by a modern general-purpose computer; conversely, the action of any computer is indeed algorithmic.

This provides a simple test of whether some suggested model or robot control system is actually algorithmic: If it can be simulated on a general purpose computer, then it is indeed algorithmic. This should make it clear that the concept of "algorithm" includes not just the action of ordinary serial computers, but parallel ones as well; and that what are referred to as "neural networks" or "connection machines" are also algorithmic, as are the operations referred to as "heuristics" and "learning" in computer systems. (Thus **Johnson, Ettinger & Hubbard**'s ARTI would certainly be algorithmic *if* he existed – despite his charmingly human display of misunderstandings of my book!) Neural networks, for example, may be constructed as special hardware, but as often as not their activity is simply simulated on a general purpose computer. I am aware that people often seem to use the phrase "algorithmic computer" in a more restrictive way than I have been doing, namely, only when the computer is programmed in a strictly "top down" way, proceeding according to some specific mathematical algorithm whose action is clearly laid down to solve a specific problem in a well-understood way. Neural networks are supposed to "learn" and modify their structure in ways that are not "preprogrammed" for the solution of specific problems, but in general ways that gradually improve their performance. The rules governing this learning and improvement, however, are themselves just algorithms (otherwise it would not be possible to simulate them on a general-purpose computer). Sometimes there may be random choices involved. Strictly, the concept of random choice would go outside what one means by algorithmic, but not usefully so. In practice, random choices are very effectively simulated in a purely algorithmic way, by the use of pseudo-random number generators. Similar remarks apply to "heuristics" (and whatever other similar terms might be or might subsequently be employed). As soon as they have been programmed on a general purpose computer, their algorithmic nature has been secured.

A number of commentators (**Boyle, Breuel, Doyle, Lutz, McDermott**) have criticized me for not addressing the question of *learning* at greater length; they appear to claim that learning is not algorithmic. In fact I did briefly refer to the question of learning in *Emperor*, in connection with neural networks (pp. 397–98), but I did not say much (and certainly nothing significant) about it. Evidently I should have done, especially now in view of such criticisms, but I had not thought that people would have taken the view that there is something new in principle here, and that some would regard a learning system as anything other than a particular type of algorithmic system. The procedures whereby the system is supposed to learn are always preprogrammed into the system right at the beginning (unless a human intervenes to modify the system at some stage – but that is clearly cheating!). A random element may be included, of course, but in practice this simply means using an (algorithmic) pseudorandom number generator. This is all just as algorithmic as before – as should be obvious from the fact that such learning systems are (or at least can be) run on an ordinary computer. Of course, a learning system is supposed to be continually influenced by its environment, so there is an external input all the time. But that is just like the Turing machine's tape, which is being continually read by the machine – and, indeed, the whole original point of the tape was that it was supposed to model the environment!

Likewise, I do not see what is to be gained by searching for the concept of "perception" in the relation between an animal and its environment (**Gilden & Lappin, Waltz & Pustejovsky**). Again, we have not got away from the Turing-machine model. The same applies also to **Varela**'s feeling that "movement" and "bodies" would make an essential difference. Actual physical movement is irrelevant to a Turing machine's action, and I find it hard to see what it has to do either with computations or consciousness. **Eagleson** refers to "intelligence" perhaps residing in "semantic levels of algorithms," but, as we have seen earlier, semantics cannot be encapsulated within an algorithm, so we are really no closer to understanding what intelligence actually is, on these terms.

Lutz seems to be saying that a human mathematician could provide provisional answers to problems all the time which are continually being improved, so it is as though we were allowed to have our Turing machine provide such provisional answers before it has actually reached a *stop* instruction. I do not really see how this buys us anything new. Such a device would need to specify the stages at which its output was allowed to be examined, and that is just like putting in a *stop* instruction at that stage instead. When it starts up again, it would just be another running of the Turing machine's action. Again, it is easy to see that all we get is something algorithmic in the old sense – clearly, because all these things can be run on an ordinary computer. Likewise, **Perlis**'s tapes, which "vary in time," do not really give us anything new if the way the tapes vary is itself well defined (and in principle computable, or computable but with random elements also). Turing's original analysis really covered all this sort of thing. The same applies to Perlis's "enlarging formalisms"; and Lutz's possibly nonterminating Turing machines are certainly algorithms, in the sense that I am using the term. That's precisely what a (possibly dud) Turing machine is. Nonterminating action must certainly be admitted. There is, after all (as **Gigerenzer** reminds us), no general (algorithmic) way of deciding whether or not a Turing machine's action does terminate!

Breuel refers to continuous systems subject to differential equations as being "not intrinsically restricted to Turing equivalent computation." In fact, a significant portion of *Emperor* was devoted to this very question. Not a great deal of a rigorous mathematical nature seems to be known about it (except for the seminal work of Pour-El & Richards, and there is the new theoretical framework of Blum et al. – both of which I refer to in *Emperor* and my *BBS* Précis). As I tried to argue in *Emperor*, there does not seem to be anything usefully nonalgorithmic in continuous action according to standard physical laws, which is one of the reasons I believe we must look to CQG for the needed usefully nonalgorithmic action. The case is no doubt arguable, but Breuel does not even mention it!

Breuel, Doyle and **Manaster-Ramer, Savitch & Zadrozny** refer to Turing's "oracle" machines and "relative computability" (and "infinite nets"). These ideas certainly do go outside what is meant by algorithmic. (An oracle is a putative device that can perform nonalgorithmic actions of the nature of giving correct yes/no answers to, for example, all propositions of type {Q} above.) The trouble is, of course, that no one knows how to build an oracle; moreover, the very possibility of constructing such a device would be denied by all those (including strong-AI supporters) who believe that we live in an algorithmic universe. Those who rest their hopes on the possibility of constructing such oracles should be pursuing my general line of endeavour and trying to locate nonalgorithmic ingredients in actual physical laws! My own proposals were, in this respect, more modest, since (despite what **McDermott** seems to suggest) I nowhere insisted that an oracle must actually be constructable but merely that nonalgorithmic ingredients be present in physical laws!

Glymour and Kelly present an argument purporting to show that even if human mental output is nonalgorithmic, we shall never be able to ascertain this fact. There seem to be two parts to the argument. First, the "we" in the above claim might actually be Turing machines; and I am certainly prepared to accept Glymour & Kelly's argument that Turing machines cannot "tell" whether or not another object is a Turing machine. The claim tells us nothing of relevance to my own arguments, however, since I am claiming that we are not Turing machines! Second, Glymour & Kelly seem to be claiming that, on some rather obscure limiting interpretation ("Δ_2 property of the Borel hierarchy"), the argument still goes through even if the "we" are not Turing machines (i.e., we are instead "noncomputable scientists"). The argument seems to boil down to something like the following: Can an oracle recognize when some other object is merely a Turing machine? An oracle would be an example of an "noncomputable scientist." We may assume that the Turing machine the oracle is examining is one that considers the truth of statements of the form [Q] one after the other, and correctly asserts *yes* or *no* in each case, except that sometimes it will get stuck and run on forever without ever giving any answer at all. In those cases the poor oracle will just have to sit it out forever (itself knowing the correct answer all the time), but at no stage will it be sure that the Turing machine isn't just a very slow oracle that hasn't yet come out with the answer!

Technically **Glymour & Kelly** are right, if I can take it that this is the kind of thing they do mean, but this is adopting an absurdly narrow view of what science can (and has been able to) achieve. We could likewise argue that the confirmation of *any* scientific theory depends on only a finite number of data points, and on a finite number of bits of information. There are infinitely many different physical theories that could fit those data, and no "scientist," in Glymour & Kelly's sense, could ever be sure that the appropriate theory had been confirmed. Yet scientific progress has undoubtedly been made – and superbly impressive progress at that. A good measure of common sense (or appropriate judgements with regard to one's statistical analysis, as **Gigerenzer** very usefully points out) is always required in the (albeit temporary) confirmation of any physical theory. Precisely the same considerations will become relevant when CGQ (or whatever turns out to the appropriate nonalgorithmic theory) finally emerges. Considerations such as those of Glymour & Kelly will not be (or at least should not be) any bar to progress!

The relevance of Gödel's theorem to cognition and AI

Several critics have argued that Gödel's theorem is *irrelevant* to the important issues of mind and AI. I completely disagree. The role of Gödel's theorem is *crucial*, though I am certainly not trying to suggest that one has to be able to appreciate Gödel's theorem to be conscious! (I can hardly be suggesting that, since in *Emperor*, I say that I believe that at least certain nonhuman animals – dolphins, chimpanzees, dogs, . . . ? – have some degree of consciousness.) Clearly one does not have to be a mathematician (or a mathematician actually doing mathematics) to be conscious. But the central role for Gödel's theorem is that it is only *here* (or with mathematical reflection principles generally) that one can demonstrate, on anything like rigorous mathematical terms, that our conscious understanding *must* be nonalgorithmic. It is, indeed remarkable that any such clear general statement about the nature of our thinking can be made at all! Without Gödel's theorem (or even with it, before the full impact of that theorem is properly appreciated) one would have to resort to much vaguer and less conclusive arguments about "semantics," "understanding," "insight," and perhaps "inspiration."

In *Emperor*, I use such arguments also – and I am criticized by **Boyle, Breuel, Eagleson, Perlis, Stanovich, Taylor, Tsotsos, Waltz & Pustejovsky** for "simply" using inconclusive and anecdotal arguments from "introspection" to support my case for nonalgorithmic action. These reviewers have simply ignored the powerful argument from Gödel's theorem, however (and others, such as Wilensky, seem to underestimate it). I never intended that the case for a nonalgorithmic ingredient to our conscious thinking should rest on anecdotal arguments of this kind. These provide only a supplement to the powerful Gödelian case, providing some additional tentative clarification as to what must really be going on. These arguments are for those who are already prepared to entertain the probability that some ingredients of our thinking are nonalgorithmic, and who are interested in gaining some insights into the possible nature of those nonalgorithmic ingredients. Like **Higginbotham**, I certainly do not believe that the nonalgorithmic quality of human thought is something restricted to the appreciation of sophisticated mathematical thinking. If it were, it would never have arisen by natural selection! It must have been useful for such things as the conception of mammoth traps and the like – or for apes perceiving the value of using tools! It is the elusive quality of *understanding* (and things related to it) that has this nonalgorithmic character – and this is what Gödel's theorem demonstrates, albeit (of necessity) in a particularly sophisticated context.

I have taken the line that consciousness is a quality that *must* be present for a nonalgorithmic (Gödelian) activity of the brain to come into play. Here I would agree with **Butterfield** and others that my case is not so strong, resting, as it does, on the fact that consciousness is (as I believe) essential for actual understanding (and hence for the appreciation of the truth of Gödel's theorem). From the different point of view of his "Chinese Room," John Searle has also used "understanding" (and related concepts such as "semantics" and "intentionality") as a fundamental quality that cannot be evoked merely by computation. These arguments do not directly address other aspects of consciousness, but I feel sure that they must ultimately have relevance for them. [See Searle: "Consciousness, Explanatory Inversion, and Cognitive Science" *BBS* 13(4) 1990.] **Butterfield, McDermott, Niall, Roskies,** and **Stanovich** take me to task for not distinguishing various *different* aspects of consciousness and for not being precise enough about what I do mean by the term. Perhaps I should have been more explicit, but I am not sure that this would have been helpful. I agree that any direct connection between the Gödelian insight and "red" or "pain" or "pride" or "hope" would be pretty hard to discern; but it seems to me that there is some sort of unified concept of consciousness, albeit a difficult one to pin down, and any small progress toward understanding some part of this concept will lead ultimately to progress toward understanding the rest of it.

McDermott objects to my "oneness" of consciousness, but I think he totally misunderstands what I mean. I know that consciousness can take multifarious forms, and also that the mind (where this term includes the unconscious mind) can indulge in many activities simultaneously. Yet, especially considering the incredible amount of unconscious activity that goes on at once in the brain, it is quite remarkable that *consciously* we feel like a single person at all. I feel sure that Freud and James would agree with me that this aspect of consciousness is quite unlike the action of a parallel computer – which was my point!

Libet raises the point that much of what is *un*conscious in mental activity could well be nonalgorithmic also (I think **Gigerenzer** and perhaps **Roeper** are making a similar point). On reflection, I think I might well agree with them. I am certainly taking the line that CQG (or whatever turns out to be the appropriate physics spanning the two quantum processes **U** and **R**) is something that can (and does) take place independently of consciousness, and I am also taking the line that this physics should turn out to be nonalgorithmic. From this I ought indeed to be inferring that important nonalgorithmic actions are taking place at much lower levels than the level of consciousness. I think, however, that consciousness itself must have some *special* role to play in relation to question of noncomputability (something to do with the level at which nonalgorithmic aesthetic truth-judgements begin to be important?), but the understanding of these matters must lie a long way beyond even CQG!

Be that as it may, AI researchers will ignore the arguments from Gödel's theorem at their peril. Mathematical thinking, although a tiny minority activity, is thinking, after all, and if that is demonstrably nonalgorithmic, then nonalgorithmic thought is shown to be possible. That is all we need from the argument. I believe that most AI people are grossly underrating the importance of a possibility of nonalgorithmic control (perhaps as exhibited by conscious action?) as a separate quality from computation – that could indeed (**Wilensky** take note) even be relevant to one's walking down the hallway!

My inadequate treatment of AI

Several critics (**Perlis, Smithers, Tsotsos, Waltz & Pustejovsky**) have chastised me for being out of date, misleading, or inadequate in my treatment of artificial intelligence. As I confessed in *Emperor*, I am aware of only a small part of the current activity in this subject, and I fully admit that a somewhat more comprehensive treatment of this activity would have been helpful. It is in the nature of this activity, however, that although there is much of it going on, I have been unable to perceive any fundamentally new principles that have emerged, of relevance to what I had to say. A survey would have amounted largely to a list of things that people in different parts of the world are doing, and that would not have been very helpful.

Many (e.g., **Perlis**) rebuke me, naturally enough, for not having referred to their own work, while **Smithers** takes me to task particularly for not referring to the "foundational" 1976 paper of Newell & Simon, where a form of "symbolic manipulation" that he claims is more relevant than Turing machine algorithmic action is described. I cannot understand why he thinks there is anything new in principle here. Newell & Simon's system, for all its virtues with regard to practicality, and so on, is nevertheless still subsumed by Turing's simple yet comprehensive original scheme. A trouble, of course, is that coming to the subject from the outside, I do not see clearly which publications AI people think I should be concentrating on. For example, no critics other than Smithers tell me that I should have referred to Newell & Simon's (1976) paper, although it receives a passing mention by **MacLennan.** I am quite prepared to accept that the paper may be important to theoretical AI, but I still do not see how it affects the arguments that I have been putting forward in any substantive way.

Tsotsos tells me that my references are out of date (though I did describe what must surely be the most impressive achievements of artificial intelligence to date: the development of such chess-playing machines as "Deep Thought," with its success in co-winning an international chess tournament in late 1988 and defeating a grand master). He thinks I would be upset if, in a description of a *physical* theory, all the references were old ones; yet the most recent reference to electromagnetic theory that I gave in my own book was to Maxwell, 1865. The important thing is to be right (like Maxwell), not necessarily to be recent! I have yet to be convinced that any substantial new advance in the theory of AI has occurred in very recent years. Chess-playing machines apart, the early things I described still seem to be the most striking achievements of AI.

Smithers maintains that I am wrong in attributing to AI workers the desire to simulate feelings of any kind, such as pleasure and pain (cf. pp. 14–17 of *Emperor*). But surely these *are* among the intentions of the supporters of *strong* AI, since *all* mental qualities are supposed to arise from computation. Weak-AI practitioners need certainly have no such intentions, and I apologize to them if they feel that they have been misrepresented. Nevertheless, the very name "artificial intelligence" implies that even the practitioners of weak AI must be attempting to simulate genuine intelligence, and actual intelligence is something that (in my view of what the word means at least, and despite what **Stanovich** says) implicitly involves some degree of awareness. Simulated pleasure and pain are certainly aspects of simulated awareness – and **Doyle** claims that a limited simulation of actual consciousness has *already* been achieved (though how he knows this is hard for me to fathom!).

CHAPTER 196

Pointers toward a new physics

Comparatively few commentators (**Butterfield, Davis, Kentridge, Lutz, MacLennan, Manaster-Ramer, Savitch & Zadrozny, Madsen, Tsotsos, Varela, Zytkow**) address the actual physical ideas I put forward in *Emperor*, despite the fact that the physics is the central theme of the work. Perhaps it is unfair to expect much discussion of such matters in a journal mainly devoted to issues of psychology, neurophysiology, and philosophy, but I am nevertheless disappointed at the overall response here because of the importance of the issues raised. **Varela** refers to my "wild speculations" (which he appears to place on a par with even the fantastical physical suggestions of Moravec 1988), whereas **Madsen** refers to my "extremely tall and particularly shaky edifice" to be accepted "on faith alone." They are unable to point to any actual flaws or omissions relevant to my chain of reasoning, however (although I consider that Madsen is being complimentary when he says that my book "raises more questions than it answers"!). Madsen asserts that I ignore Mott's work on α-particles (citing no reference – but presumably he means a paper I actually refer to in my reference list in *Emperor* – Mott 1983), appearing to claim that Mott showed that large quantum systems will necessarily behave like classical ones – though Mott actually did nothing of the sort; nor was he trying to. Madsen is upset that I give short shrift to the still somewhat fashionable inflationary models (though my placing them in the TENTATIVE category is surely unexceptionable). Such models usually do depend on GUT theories; but the fact that some may not is neither here nor there. They do not in any way alter the arguments that I give in my book (see Penrose 1989 for a critical discussion of the relevant points). Madsen closes with some very odd ideas about "particles . . . using Turing's methods to compute their trajectories in phase space," although such ideas certainly played no role in my own discussions – nor would a theory of the (deliberately?) absurd type suggested by **McDermott** at the beginning of his critique.

Various criticisms refer disparagingly to my search for ideas that might lead to an improved quantum mechanical theory that would be more satisfactory with regard to the "measurement problem" ("vague thought," says **McDermott**, "purest speculation" according to **Madsen** – though these critics offer no counterarguments, apparently forming their judgements on the basis of my own apparently over-modest reference to my suggestions being "only a germ of an idea," p. 371 of *Emperor*). **Tsotsos** and **Zytkow** appear to have totally misunderstood my "one-graviton criterion," incorrectly thinking that it refers to effects at 10^{-33} cm or else to *real* graviton emission or absorption. No doubt I was not clear enough: Only virtual, or *longitudinal* gravitons are needed (cf. Note 5 on p. 372). These can occur at a much tinier level of energy-difference than real gravitons, and the relevant scales of distance can be as large as you like (e.g., ten metres or so, as in the Aspect experiment; cf. Aspect & Grangier 1986). Moreover, I am referring to *all* wavefunction collapse, not just the "very special circumstances of quantum gravity" **Zytkow**; and that would be taking place in the brain all the time.

I should have thought I had made it clear that in my view, wavefunction collapse (**R**) occurs spontaneously, quite independently of the presence or absence of conscious observers. I do *not* hold to the view that it is the "interaction between minds and the rest of the universe" (**MacLennan**) that leads to **R**; nor do I believe in a "subjective view of the state vector involving conscious observers" (**Boyle**). Moreover, I did not say that I believe the brain to be a quantum computer, in the presently understood sense – though there may well be aspects of that conception in brain function; and despite what MacLennan says in a footnote, there are very fundamental distinctions between quantum and classical linearity.

I also find it remarkable how many AI people appear to regard the many-worlds interpretation as "obviously true" (**MacLennan, McDermott**), despite its numerous problems and (deserved) unpopularity in the physics community as a whole – and the fact that it does not (in the absence of further nonstandard ingredients) allow one to derive the correct quantum probabilities. (The "observer being part of the physical universe," which I believe in as much as does MacLennan, is no explanation of why a many-worlds universe looks like one world!) I trust that their reasons for believing in the validity of the AI programme are more soundly based.

Kentridge's remarks about the possible effects of quantum processes in computation are well thought through and helpful (if perhaps somewhat inconclusive), but I do not think that **Mortensen** or **Roeper** have yet grasped the point that quantum mechanics is not just a "fuzziness" in our classical descriptions, but a theory of the utmost precision. Quantum processes can lead to effects that are not possible to achieve by classical means. I should point out to **Lutz**, however, that Deutsch's (1985) analysis of quantum computation is that it does not, in itself, lead to nonalgorithmic behaviour – which is one reason I believe that more (e.g., CQG) is needed.

Butterfield refers to my speculations, in relation to the timing of consciousness, that putative causality-violating effects might sometimes *not* lead to inconsistency (because of the timelessness of the Platonic world). His criticisms are indeed pertinent, and I agree with him that any too blatant causality violation by a "Platonic realization" could lead to serious consistency problems. My remarks were supposed to be exploratory only, however, and were intended to indicate that there might be more scope for temporal nonlocality in consciousness than one might otherwise have thought. Whether or not any kind of a consistent scheme can be worked out remains to be seen.

I had in mind, with these considerations, the curious time delays involved in the experiments of Kornhuber and of Libet (Libet et al. 1979). I am grateful for **Libet**'s reassurance that I had not misrepresented him, and also for his comments in relation to his own more recent work in following up the Kornhuber-type experiments. (I had become vaguely aware of these more recent experiments just before I completed my book, but I had not had the opportunity to follow them up.) It seems to me that there are still some very puzzling aspects of the combined force of all these important experiments (despite what **McDermott** says), but I think I had better reserve my judgements until I have had the opportunity to examine these newer results a little more closely.

Perhaps I may comment, at this point, that I am also grateful to **Eccles** for his remarks, and that I am very much in agreement with his belief that scientific methods – and particularly neurophysiological experiments, such as those he and **Libet** have pioneered, in addition to a deeper probing of physical laws – must hold the key to our eventual understanding of the mind/brain issue. I shall try to follow up his comment concerning the "elimination" of the brainstem centre theory of consciousness (and also his recent work referred to in his commentary). Experimental verification of all these ideas (my own included) is always the final test. The one-graviton suggestion for the onset of **R**, for example, should certainly be experimentally testable, and certain tentative ideas along these lines have already been mooted.

If the physical suggestions I have been putting forward can be formulated as a fully fledged nonalgorithmic CQG, then that theory (including its very nonalgorithmic nature – in spite of **Glymour & Kelly**'s remarks) will eventually also be put to the experimental test. If nonalgorithmically behaving objects are ever shown to exist in nature, then that will alter the whole focus of what we actually mean by a "machine" or, rather, by an analogue machine. (According to the terminology of **Chalmers** and **Manaster-Ramer, Savitch & Zadrozny**, the Church(-Turing) thesis would then be false – though I prefer the original terminology according to which that thesis refers to a *mathematical* idea, now almost universally accepted as true.)

Platonism, mysticism, and dualism

Several commentators (**Garnham, Higginbotham, MacLennan, Varela, Waltz & Pustejovsky**) express discomfort (or worse, **Madsen**) with the idea of mathematical *Platonism* – some seeming to identify it with mysticism. Yet the evidence from the majority of mathematicians (who appear to be Platonists of some degree, cf. Davis & Hersch 1982) should not be ignored. It is the mathematicians, after all, who know most about their subject. Though Platonism does have certain difficulties (mainly to do with deciding which enormous sets should be considered actually to exist, or the problem of deciding when to accept a "proof" as a proof), the various alternative viewpoints (formalism, intuitionism, finitism) all have much more severe difficulties (cf. **Garnham**). **Higginbotham** writes that the mathematicians have "had to learn to live with incomplete theories," but it is only the formalists whose mathematics has been rendered incomplete by Gödel. The Platonists are not so constrained. Moreover, the evidence for a Platonic existence for such things as the Mandelbrot set seems to me to be overwhelming; nor does a pragmatist view (as expressed by **Garnham**) come to terms with the extraordinary agreement between mathematicians as to whether or not a mathematical result is true, or with the very fact that they settle their issues by abstract reasoning. It would have been helpful if some of the opponents of Platonism had provided more convincing arguments against such powerful cases for a Platonist view. The facts that certain mathematical concepts may take many years to take shape and that nonstandard analysis can coexist alongside standard analysis (just as nonEuclidean geometry can coexist with Euclidean geometry) are, despite what **MacLennan** seems to be claiming, perfectly consistent with Platonism.

As for mysticism, I have never myself made that association. Platonism is, to me, totally consistent with (and even a concomitant of) a completely scientific viewpoint. But it if is insisted that Platonism entails mysticism, then I shall insist that it is the only form of mysticism I can accept. What about dualism? **Waltz & Pustejovsky**, not to mention **MacLennan**, seem to be insisting that I am a dualist, presumably because of my Platonistic opinions. If I am, then it can only be in the sense that the strong-AI people are also dualists – because they believe that Platonically existing algorithms provide the substance of awareness. (**Higginbotham** has trouble with the apparently temporal aspect of my argument: "Strong-AI supporters must be Platonists, because if algorithms exist only in minds, pre-existing minds are needed for algorithms and pre-existing algorithms for minds!" But I intended this argument to be taken logically, not temporally, and then the circularity is surely clear.) If my Platonism (with regard to the nonalgorithmic procedures I believe must underlie consciousness) makes me a dualist, then, again, that would be the only form of dualism I could accept. Unlike the AI supporters, however, I have not made the claim that it is the mere enaction (or existence) of some mathematical process that evokes consciousness.

Perlis writes about the "dynamic action of algorithms" as being what AI people are concerned with – and presumably what strong-AI people would claim evokes awareness – rather than the mere static existence of an algorithm. This is totally unclear, however. "Dynamic action" presumably means that what counts is the moving around of bits of matter in accordance with the algorithm. The (strong) AI position seems to be that it does not make any difference what the bits of matter that are being moved around actually are. If the nature of the matter is irrelevant, why is the matter itself relevant? What about showing a film of the pages in a book in which the working of the algorithm is written out – or just running one's finger (or a little window in a piece of paper) down the page? The viewpoint is unclear at best.

ACKNOWLEDGMENT

I am grateful to Angus MacIntyre for very useful comments and for his support; also thanks go to Rob Baston for his patience and use of his E-mail address.

NOTES

1. Ironically, a truly committed formalist would accept G(F) *only* if it is false! For only then is it actually a theorem of F.

2. Examples of such Q(n)s would be: "6n is divisible by 3"; "if n is prime, then 2n + 1 is prime"; "n is the sum of four squares"; "2n + 4 is the sum of two primes". For the corresponding universally quantified proposition [Q], we have, in each case: The first is obviously true, the second false, the third nonobviously true (by a theorem of Lagrange), and the fourth (Goldbach's conjecture) unknown. Although I suppose that a case can be made (according to a fairly extreme form of intuitionism) that it is not even meaningful to say whether statements of the form [Q] are either true or false unless they have been "demonstrated" to be, none of my critics have followed that unreason-

able (to my mind) stance (except for a brief allusion to it by **Mortensen**) and I shall not pursue it.

3. If we allowed an infinite set of theorems here, we would need some procedure for algorithmically generating them. The argument given in the text still applies (cf. pp. 109–10 of *Emperor*). **Chalmers** please note.

4. In fact, one can strengthen the argument I gave, since it applies to any "stopping-tester" $H(m;n)$ which is allowed even to make mistakes or to run on forever – provided that it does not ever assert that $T_m(n)$ runs on forever when in fact $T_m(n)$ stops (the only directly falsifiable case!). (Thus, $H(m;n) = 0, 1$ or $\Box$, if $T_m(n) = \Box$; and $H(m;n) = 1$ or $\Box$, if $T_m \neq \Box$. Here, in the notation of my book, '$\Box$' means "does not stop" and $H = 1$ asserts that the action of $T_m(n)$ stops while $H = 0$ asserts that it does not.) We construct k so that $1 + T_n(n) \times H(n;n) = T_k(n)$ and consider the action $T_k(k)$. We can see that this action does not stop, but H either gives no answer or gives the wrong answer. It would not be too hard to construct a completely explicit Turing machine T_q that gives k in terms of the Turing number of H (with q probably less than half as long again as the number u that I gave in my book for T_u to be a universal Turing machine). One might think that since there is an algorithm for producing the number k that defeats H, we could put the whole (Turing-)Gödelization process on a machine, but this is not so, for the same reason as in the argument in the main text.

5. According to a *reflection principle*, an understanding of the *meaning* underlying the rules of a formal system F can provide new mathematical truths not accessible within F (see p. 110 of *Emperor* or, more extensively but with a somewhat different emphasis, Rucker 1984).

References

Agree, P. (1988) The dynamic structure of everyday life. Artificial Intelligence Laboratory Technical Report 1085, Massachusetts Institute of Technology. [FJV]

Albus, James S. (1981) *Brains, behavior, and robotics*. BYTE Books. [RE]

Allport, D. A. (1980) Attention and performance. In: *Cognitive psychology: New directions*, ed. G. Claxton. Routledge & Kegan Paul. [KES]

Anscombe, G. E. M. (1981). *Metaphysics and the philosophy of mind*. Basil Blackwell. [KKN]

Armstrong, D. M. & Malcom, N. (1984) *Consciousness & causality*. Basil Blackwell. [KES]

Arnhart, L. (in press). Aristotle, chimpanzees, and other political animals. *Social Science Information*.

Aspect, A. & Grangier, P. (1986) Experiments on Einstein-Poddsky-Rosen-type correlations with pairs of visible photons. In: *Quantum concepts in space and time*, ed. R. Penrose & C. J. Isham. Oxford University Press. [rRP]

Baerends, G. P. & Kruijt, J. P. (1973) Stimulus selection. In: *Constraints on Learning*, ed. R. A. Hinde & J. Stevenson-Hinde. Academic Press. [DH]

Barrow, J. O. & Tieler, F. J. (1986) *The anthropic cosmological principle*. Oxford University Press. [MSM]

Bell, John L. (1986) A new approach to quantum logic. *British Journal of the Philosophy of Science* 37:83–99 [RE]

Bell, J. S. (1987) *Speakable and unspeakable in quantum mechanics*. Cambridge University Press. [aRP]

Benacerraf, P. (1967) God, the Devil, and Gödel. *The Monist* 51(1):9–32. [JH, AR]

Berliner, H. (1988) New hitech computer chess success. *AI Magazine* 9(2):133. [TS]

Bishop, E. (1967) *Foundations of constructive analysis*. McGraw-Hill. [BM]

Blum, L., Shub, M. & Smale, S. (1989) On a theory of computation and complexity over the real numbers: NP completeness, recursive functions and universal machines. *Bulletin of the American Mathematics Society* 21:1–46. [aRP]

Bohm, D. (1980) *Wholeness and the implicate order*. Routledge & Kegan Paul. [DH]

Brooks, R. (1986a) A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2, April, 14–23. [DW]

(1986b) Achieving artificial intelligence through building robots. Artificial Intelligence Laboratory Memo 899. Massachusetts Institute of Technology. [FJV]

Caines, P., Greiner, R. & Wang, S. (1989) Classical and logic-based dynamic observers for finite automata. Technical Report, McGill University, May 12. Draft copy. [RE]

Cariani, P. (1989) On the design of devices with emergent semantic functions, unpublished Ph.D. thesis. State University of New York at Binghamton. [DW]

Casti, J. L. (1989) *Alternate realities: Mathematical models of nature and man*. Wiley. [DLC]

Chisholm, R. M. (1957) *Perceiving*. Cornell University Press. [KKN]

Chomsky, N. (1986) *Knowledge of language*. Praeger. [TR]

Churchland, P. M. (1979) *Scientific realism and the plasticity of mind*. Cambridge University Press. [KES]

(1988) *Matter and consciousness*, 2nd ed. MIT Press. [KES]

Churchland, P. M. & Churchland, P. S. (1990) Could a machine think? *Scientific American* 262(1):32–7. [KES]

Churchland, P. S. (1983) Consciousness: The transmutation of a concept. *Pacific Philosophical Quarterly* 64:80–95. [KES]

(1986) *Neurophilosophy: Toward a unified science of the mind/brain*. MIT Press. [KES]

Cutting, J. E. (1986) *Perception with an eye to motion*. MIT Press. [KKN]

Davis, M. (1977) A relativity principle in quantum mechanics. *International Journal of Theoretical Physics* 16:867–74. [MD]

Davis, M., Matijasevic, Y. & Robinson, J. (1976) Hilbert's tenth problem: Diophantine equations; positive aspects of a negative solution. *Proceedings of Symposia in Pure Mathematics* 28:323–78. [MD]

Davis, P. J. & Hersh, R. (1982) *The mathematical experience*. Harvester Press. [rRP]

(1986) *The blind watchmaker*. Longman Scientific and Technical. [DH]

Deecke, L., Grozinger, B. & Kornhuber, H. H. (1976) Voluntary finger movements in man: Cerebral potentials and theory. *Biological Cybernetics* 23:99–119. [BL]

Delbrueck, M. (1986) *Mind from matter?* Blackwell Scientific Publications. [DH]

Dennett, D. C. (1965) What numbers could not be. *Philosophical Review* 74:47–73. [AR]

(1969) *Content and consciousness*. Routledge and Kegan Paul. [KES]

(1978a) The abilities of men and machines. In: *Brainstorms*. MIT Press. [AR]

(1978b). *Brainstorms*. MIT Press. [KES]

(1987) *The international stance*. MIT Press. [DP, KES]

(1988) When philosophers encounter artificial intelligence. *Daedalus* 117:284–95. [KES]

(1989) Murmurs in the cathedral. *Times Literary Supplement*, The London Times, Sept 29–Oct. 5 [DCD, DH]

Deutsch, D. (1985) Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London* A400:97–117. [arRP, RL]

Dewitt, B. S. (1973) *The many-worlds interpretation of quantum mechanics*. Princeton University Press. [aRP, DM]

Dewitt, B. S. & Graham, R. D. (eds.) (1973) *The many-worlds interpretation of quantum mechanics*. Princeton University Press. [aRP, DM]

Dirac, P. A. M. (1938) Classical theory of radiating electrons. *Proceedings of the Royal Society of London* A167:148. [aRP]

Dreyfus, H. (1979) *What computers can't do*. Harper & Row. [FJV]

Eccles, J. C. (1986) Do mental events cause neural events analogously to the probability fields of quantum mechanics? *Proceedings of the Royal Society of London* B 227:411–28. [DH, JCE]

(1990) A unitary hypothesis of mind-brain interaction in the cerebral cortex. Proceedings of the Royal society of London B. [JCE]

Einstein, A., Podolsky, P. & Rosen, N. (1935) Can quantum-mechanical descriptions of physical reality be considered complete? *Physics Review* 47:777–80.

Elgot-Drapkin, J. (1988) Step-logic: Reasoning situated in time. Ph.D dissertation, Computer Science Department, University of Maryland. [DP]

Epstein, R., Kirshnit, C. E., Lanza, R. P. & Rubin, L. C. (1984) "Insight" in the pigeon: Antecedents and determinants of an intelligent performance. *Nature* 308:61–62. [DH]

Everett, H. (1957) "Relative state" formulation of quantum mechanics. *Review of Modern Physics* 29:454–62. [aRP, BM]

Fodor, J. A. & Pylyshyn, Z. W. (1981) How direct is visual perception? Some reflections on Gibson's "ecological approach." *Cognition* 9:139–96. [KKN]

Fodor, J. (1983) *The modularity of mind*. MIT Press. [DW]

Fontana, W. & Schuster, P. (1987) A computer model of evolutionary optimization. *Biophysical Chemistry* 26:123–47. [DH]

Geroch, R. & J. B. Hartle (1986) Computability and Physical Theories. *Foundations of Physics* 16:533–50. [AM-R]

Ghirardi, G. C., Rimini, A. & Weber, T. (1986) Unified dynamics for microscopic and macroscopic systems. *Physical Review* D34:470. [aRP]

Gibson, J. J. (1979) *The ecological approach to visual perception*. Houghton Mifflin. [KKN, DLG, FJV]

Gibson, W. (1984) *Neuromancer*. Gollancz. [CM]

Gigerenzer, G. & Murray, D. J. (1987) Cognition as intuitive statistics. Erlbaum. [GG]

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Kruger, L. (1989) *The empire of chance. How probability changed science and everyday life*. Cambridge University Press. [GG]

Ginsberg, M. (1987) *Readings in non-monotonic reasoning*. Morgan Kaufmann. [DP]

Gödel, K. (1964) What is Cantor's continuum problem? In: *Philosophy of mathematics*, ed. P. Benacerraf & H. Putnam. Prentice-Hall. [KKN]

(1986/1990) *Kurt Gödel, collected works*, vol. I (orig. pub. 1929–1936) & vol. II (orig. pub. 1938–1974), ed. S. Feferman, J. W. Dawson Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, J. van Heijenoort. Oxford University Press. [rRP]

Goldberg, D. S. (1989) *Genetic algorithms*. Addison-Wesley [DH]

Halmos, P. (1974) *Measure theory*. Springer. [CG]

Harnad, S. (1989) Minds, machines, and Searle. *Journal of Experimental and Theoretical AI* 1(1):5–25. [TS]

(1990) The symbol grounding problem. *Physica D* 42:335–46. [DW, TS]

Harth, E. (1982) *Windows on the mind*. Harvester Press. [BL]

Haugeland, J. (1985) *Artificial intelligence: The very idea*. MIT Press. [DM]

Hebb, D. O. (1954) The problem of consciousness and introspection. In: *Brain mechanisms and consciousness*, ed. J. F. Delafresnaye. Blackwell. [aRP]

Hilbert, D. & Ackermann, W. (1928) *Grundzüge der Theoretischen Logik*. Springer. [MD]

Hillis, D. (1988) Intelligence as an emergent behavior, or, The songs of Eden. *Daedalus* 117(1):175–90. [DW]

Hinman, P. (1978) *Recursion-theoretic hierarchies*. Springer [CG]

Hodges, A. P. (1983) Alan Turing: *The enigma*. Simon and Schuster. [rRP]

Hofstadter, D. R. (1979) *Gödel, Escher, Bach: An eternal golden braid*. Basic Books. [RWK]

(1985) *Metamagical themas*. Basic Books. [KES]

Holland, J. (1975) *Adaptation in natural and artificial systems*. University of Michigan Press. [DH]

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America* 79:2554–58. [RWK]

Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly*. [JB]

Johnson-Laird, P. N. (1983) *Mental models*. Cambridge University Press. [AG]

Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press. [RWK]

Kelly, K. (1990) Characterizations of inductive inference on arbitrary sets of sequences. Laboratory for Computational Linguistics, Department of Philosophy, Carnegie Mellon University. [CG]

Kihlstrom, J. (1987) The cognitive unconscious. *Science* 237:1445–52. [KES]

Kleene, S. C. (1950) *Introduction to metamethametics*. Van Nostrand. [AM-R]

Lakatos, I. (1976) *Proofs and refutations: The logic of mathematical discovery*. Cambridge University Press. [BM]

(1976) *Proofs and refutations*. Cambridge University Press. [CM]

Lewis, D. (1969) Lucas against mechanism. *Philosophy* 44:231–33. [JB]

(1979) Lucas against mechanism II. *Canadian Journal of Philosophy* 9:373–76. [JB]

(1990) What experience teaches. In: *Mind and cognition: A reader*, ed. W. Lycan. Cambridge University Press. [JB]

Libet, B. (1965) Cortical activation in conscious and unconscious experience. *Perspectives in Biology and Medicine* 9:77–86. [BL]

(1980) Mental phenomena and behavior. *Behavioral and Brain Sciences* 3:434. [BL]

(1982) Brain stimulation in the study of neuronal functions for conscious sensory experiences. *Human Neurobiology* 1:235–42. [BL]

(1985) Unconscious cerebral intiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8:529–66. [BL]

(1987) Consciousness: Conscious subjective experience. In: *Encyclopedia of neuroscience*, vol. I, ed. G. Adelman. Birkhauser. [BL]

(1989) Conscious subjective experience vs. unconscious mental functions: A theory of the cerebral processes involved. In: *Models of brain function*, ed. R. M. J. Cotterill. Cambridge University Press. [BL]

Libet, B., Gleason, C. A., Wright, E. W. Jr. & Pearl, D. K. (1983) Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential); the unconscious initiation of a freely voluntary act. *Brain* 106:623–42. [BL]

Libet, B., Wright, E. W. Jr., Feinstein, B. & Pearl, D. K. (1979) Subjective referral of the timing of a conscious sensory experience. A functional role for the somatosensory specific projection system in man. *Brain* 102:193–224. [rRP, BL]

Libet, B., Wright, E. W. Jr., & Gleason, C. A. (1982) Readiness potentials preceding unrestricted "spontanious" vs. pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology* 54:322–35. [BL]

Lucas, J. R. (1961) Minds, machines and Gödel. *Philosophy* 36:112–27. [arRP, JB, RWK, DJC]

Lyons, W. (1986) *The disappearance of introspection*. MIT Press. [KES]

Mace, W. M. (1977) James J. Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In: *Perceiving, acting, and knowing*, ed. R. Shaw and J. Bransford. Lawrence Erlbaum Associates. [DLG]

MacLennan, B. J. (1987) Technology-independent design of neurocomputers: The universal field computer. *Proceedings of the IEEE First Annual International Conference on Neural Networks* III:39–49. [BM]

Margenau, H. (1984) *The miracle of existence*. Ox Bow Press. [DH]

Marr, D. (1977) Artificial intelligence – a personal view. *Artificial Intelligence* 9:37–48. [DH]

Martin, R. (ed.) (1984) *Recent essays on truth and the liar paradox*. Oxford University Press. [JH]

Maturana, H. & Varela, F. (1987) *The tree of knowledge: The biological roots of human understanding*. New Science Library. [FJV]

McAlister, L. L. (1976) *The philosophy of Brentano*. Duckworth. [KKN]

McCammon, J. A. & Harvey, S. C. (1987) *Dynamics of proteins and nucleic acids*. Cambridge University Press. [DH]

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence*, ed. M. Ringle. Humanities Press [DP]

McClelland, J. L. & Rumelhart, D. E. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press. [KES]

McGinn, C. (1987) Could a machine be conscious. In: *Mindwaves*, ed. C. Blakemore & S. Greenfield. Blackwells. [DH]

Mitchell, M. & Hofstadter, D. R. (in press) The emergence of understanding in a computer model of concepts and analogy-making. *Physica D*. [DJC]

Minsky, M. (1986) *Society of mind*. Simon and Schuster. [DW, DM, KES]

Monod, J. (1972) *Chance and necessity*. Random Paperback Knopf. [DH]

Moravec, H. (1988) *Mind children: The future of robot and human intelligence*. Harvard University Press. [rRP, FJV]

Mortensen, C. (1989) Mental images. In: *Computers, brains, and minds*, ed. P. Slezack & W. Albury, Kluwer. [CM]

Mott, N. F. (1983) The wave mechanics of α-ray tracks. In: *Quantum theory and measurement*, ed. J. A. Wheeler & W. H. Zurek. Princeton University Press A126:79–84 (orig. pub. 1929, Proceedings of the Royal Society of London). [rRP]

Nagel, E. & Newman, J. R. (1958) *Gödel's proof*. Routledge & Kegan Paul Ltd. [rRP]

Nagel, T. (1979) *Mortal questions*. Cambridge University Press. [DH]

Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [RE]

(1982) The knowledge level. *Artificial Intelligence* 18:87–127. [RE]

Newell, A. & Simon, H. A. (1976) Computer science as empirical enquiry: Symbols and search. *Communications of the ACM* 19:113–26. [rRP, BM, TS]

Nisbett, L. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall. [KES]

Nisbett, L. & Wilson, T. (1977) Telling more than we know: Verbal reports on mental processes. *Psychological Review* 84:231–59. [KES]

Pearle, P. (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. *Physical Review* A39:2277–89. [aRP]

Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.

Penrose, R. (1989) Difficulties with inflationary cosmology. Proceedings of the 14th Texas Symposium on Relativistic Astrophysics, ed. E. Fenves. New York Academy of Science. [rRP]

(1990) Matter over mind. *The New York Review of Books*, Feb. 1, 37(1):3–5. [FJV]

Perlis, D. (1987) How can a program mean? *Proceedings, International Joint Conference on Artificial Intelligence*, Milan. Morgan Kaufmann Publishers. [DP]

(1989) Some brief essays on mind. Technical Report 302, Computer Science Department, University of Rochester. [DP]

Popper, K. R. & Eccles, J. C. (1977) *The self and its brain*. Springer Verlag. [JCE]

Posner, M. I. (1989). *Foundations of cognitive science*. MIT Press. [KES]

Pour-El, M. B. and Richards, I. (1981) The wave equation with computable initial data such that its unique solution is not computable. *Advances in Mathematics* 39:215–239. [aRP]

Putnam, H. (1960) Minds and machines. In: *Dimensions of mind*, ed. S. Hook. New York University Press. [CM]

(1988) *Representation and reality*. MIT Press. [KKN]

Pylyshyn, Z. W. (1980) The "causal power" of machines. *Behavioral and Brain Sciences* 3:442–44. [KES]

(1984) *Computation and cognition*. MIT Press. [RE]

Rayner, K. & Pollatsek, A. (1989) *The psychology of reading*. Prentice-Hall. [KES]

Robinson, A. (1966) *Non-standard analysis*. North-Holland. [BM]

Rogers, H. (1968) *Theory of recursive functions and effective computability*. McGraw-Hill. [CG]

Rollman, G. B. & Nachmias, J. (1972) Simultaneous detection and recognition of chromatic flashes. *Perception & Psychophysics* 12:309–14. [KES]

Rorty, R. (1979) *Philosophy and the mirror of nature*. Princeton University Press. [KES]

Rosen, R. (1978) *Fundamentals of measurement and the representation of natural systems*. North Holland. [DW]

(1987) On the scope of syntactics in mathematics and science: The machine metaphor. In: *Real brains, artificial minds*, ed. J. Casti & A. Karlqvist. North Holland. [DW]

Rucker, R. (1984) *Infinity and the mind: The science and philosophy of the infinite*. Paladin Books/Granada Publishing Ltd. [rRP]

Ryle, G. (1949) *The concept of mind*. Barnes & Noble. [KES]

Saltzman, E. (1979) Levels of sensorimotor representation. *Journal of Mathematical Psychology* 20:91–163. [RE]

Schank, R. & Abelson, R. (1977) *Scripts, plans, goals, and understanding*. Erlbaum. [JKT]

Schrödinger, E. (1935) Die gegenwärtige situation in der Quantenmechanik. Naturwissenschaftenp 23:807–12, 823–28, 844–49. [aRP]

Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–50. [aRP, DP, DM, RW]

(1983) *Intentionality*. Cambridge University Press. [KKN]

(1984) *Minds, brains and science*. Harvard University Press. [GC, FJV]

(1990) Is the brain's mind a computer program? *Scientific American* 262:26–31. [DW]

Seidenberg, M. S. & McClelland, J. L. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96:523–68. [KES]

Sejnowski, T. J. & Rosenberg, C. R. (1988) Learning and representation in connectionist models. In: *Perspectives in memory*, ed. M. S. Gazzaniga. MIT Press. [KES]

Shankar, N. (1986) Proof checking metamathematics. Ph.D. dissertation, Department of Computer Science, University of Texas at Austin. [RW]

Sherrington, C. S. (1940) *Man on his nature*. Cambridge University Press. [JCE]

Simon, H. (1973) The organization of complex systems. In: *Hierarchy theory*, ed. H. H. Pattee. George Braziller. [RE]

(1981) *Sciences of the artificial*. MIT Press. [DW]

Smith, P. & Jones, O. R. (1986) *The philosophy of mind*. Cambridge University Press. [KES]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [DJC]

Springer, S. & Deutsch, G. (1985) *Left brain, right brain*. W. H. Freeman. [KES]

Stanfill, C. & Waltz, D. (1986) Toward memory-based reasoning. *Communications of the ACM* 29(12):1213–28. [DW]

Stanovich, K. E. (1986) *How to think straight about psychology*. Scott, Foresman. [KES]

(1989) Implicit philosophies of mind: The dualism scale and its relationships with religiosity and belief in extrasensory perception. *Journal of Psychology* 123:5–23. [KES]

Sternberg, R. J. & Detterman, D. K. (1986) *What is intelligence?* Ablex. [KES]

Stich, S. (1983) *From folk psychology to cognitive science*. MIT Press. [KES]

Sumida, B. H., Houston, A. I., McNamara, J. M. & Hamilton, W. D. (in press) Genetic algorithms and evolution. *Journal of Theoretical Biology*. [DH]

Tranel, D. & Damasio, A. (1985) Knowledge without awareness: An autonomic index of facial recognition by prospagnosics. *Science* 228:1453–54. [KES]

Tsotsos, J. K. (in press) Analyzing vision at the complexity level. *Behavioral and Brain Sciences* 13(3). [JKT]

Waltz, D. (1988) The prospects for truly intelligent systems. *Daedalus* 117(1):191–212. [DW]

Whiteman, M. (1967). *Philosophy of space and time*. Allen & Unwin. [DH]

Wigner, E. P. (1961) Remarks on the mind-body problem. In: *The scientist speculates*, ed. I. J. Good. Heinemann. [aRP]

Wilkes, K. V. (1984) Is consciousness important? *British Journal of Philosophy of Science* 35:223–43. [KES]

(1988) *Real people: Personal identity without thought experiments*. Oxford University Press. [KES]

Winograd, T. & Flores, F. (1987) *Understanding computers and cognition*. Addison-Wesley. [FJV]

Wittgenstein, L. (1967) *Remarks on the foundations of mathematics*. Blackwell. [AG]

Wolf, F. A. (1989) On the quantum physical theory of subjective antedating. *Journal of Theoretical Biology* 136:13–19. [DH]

Yolton, J.W. (1984) *Perceptual acquaintance from Descartes to Reid*. University of Minnesota Press. [KKN]

Zohar, D. (1990) *The quantum self*. Bloomsbury. [DH]

*Abstract: *The emperor's new mind* (hereafter *Emperor*) is an attempt to put forward a scientific alternative to the viewpoint of "strong AI," according to which mental activity is merely the acting out of some algorithmic procedure. John Searle and other thinkers have likewise argued that mere calculation does not, of itself, evoke conscious mental attributes, such as understanding or intentionality, but they are still prepared to accept the action the brain, like that of any other physical object, could in principle be simulated by a computer. In *Emperor* I go further than this and suggest that the outward manifestations of conscious mental activity cannot even be properly simulated by calculation. To support this view, I use various arguments to show that the results of mathematical insight, in particular, do not seem to be obtained algorithmically. The main thrust of this work, however, is to present an overview of the present state of physical understanding and to show that an important gap exists at the point where quantum and classical physics meet, as well as to speculate on how the conscious brain might be taking advantage of whatever new physics is needed to fill this gap to achieve its nonalgorithmic effects.

*From page 643.

197

TWISTOR THEORY AFTER 25 YEARS – ITS PHYSICAL STATUS AND PROSPECTS

Twistors in Mathematics and Physics Eds T. N. Bailey, R. J. Baston, 1–29. London Mathematical Society Lecture Note Series, 156, Cambridge University Press, 1990

This article outlines my view of the status of twistor theory as of around 1990, indicating what it had achieved and what it had still failed to achieve at that stage, and how I had envisaged that it might perhaps proceed in the future. Some noteworthy points are as follows: (1) on p. [3], the proposal of Hodges for regularizing twistor contour integrals through the replacement of $Z^\alpha W_\alpha$ by $Z^\alpha W_\alpha + k$ is outlined. This should be $Z^\alpha W_\alpha - k$, but in any case his proposal has undergone significant changes in recent years, with powerful and fascinating results (to be published); (2) comments are made here on pp. [10] and [11] making reference to the issue that stimulated our interest extended Regge trajectories (Chapters 95 and 194); (3) the idea expressed towards the end of p.[17] that the Weyl curvature hypothesis might imply an essentially *flat-space* twistor structure at the big bang is something that I would now disagree with, yet there should be a smooth and manageable twistor space there, in accordance with ideas I am putting forward in *Cycles of Time* (Bodley Head, 2010).

Twistor theory after 25 years – its physical status and prospects. *Twistors in Mathematics and Physics* Eds T. N. Bailey, R. J. Baston, 1–29. London Mathematical Society Lecture Note Series, 156, Cambridge University Press, 1990

Twistor Theory After 25 Years — its Physical Status and Prospects

R. Penrose

Introduction

The primary objective of twistor theory originally was—and still is—to find a deeper route to the workings of Nature; so the theory should provide a mathematical framework with sufficient power and scope to help us towards resolving some of the most obstinate problems of current physical theory. Such problems must ultimately include: (1) removing the infinities of quantum field theory, (2) ascertaining the nature and origin of symmetry and asymmetry in the classification of particles and in physical interactions, (3) deriving, from some fundamental principle, the strengths of coupling constants and the masses of particles, (4) finding a quantum gravity theory capable of satisfactorily addressing the issues raised by space-time singularities and the structure of space-time in the small, (5) constructing a picture that makes sense of the puzzling non-locality and conceptual peculiarities inherent in the process of quantum measurement. Does twistor theory have anything of significance to contribute concerning these matters? Might it at least point us in some appropriate directions?

I shall comment on these issues individually in a moment. But as things stand, it must be said that the successes of twistor theory to date have been almost entirely in applications within *mathematics*, rather than in furthering our understanding of the nature of the physical world. I would think of twistor theory's physical role, so far, as being something perhaps resembling that of the Hamiltonian formalism. That formalism provided a change in the framework for classical Newtonian theory rather than a change in Newtonian theory itself. The Hamiltonian scheme (at least Hamilton's own part in its development) was motivated very much by a physical analogy between the behaviour of particles and of waves; but it was not until the advent of quantum physics that a change in physical theory was put forward—indeed one in which particles and waves became actually the same thing, rather than being merely analogous. When the mathematics for a quantum theory was required, Hamiltonian formalism was in place and provided the ideal vehicle, ready to accommodate the essential changes that were needed, in order that physical theory could be transported from classical to quantum. The ambitious role set out for twistor theory, then, is that, likewise, when enough of its mathematics has been developed, that theory also will be in place and, with relatively minor changes, will turn out to be just what is required for a much needed new physics.

In this article, I shall be concerned primarily with physical issues, and how I feel that twistor theory stands, or ought to stand, with regard to them. The *mathematical* applications of the theory are well covered by articles by other authors in this volume, and some of these applications have proved to be unexpectedly fruitful. With regard to *physical* applications and aside from developments connected with the fundamental issues referred to above, which I discuss in a moment), there has one noteworthy and *unanticipated* success: the concept of *quasi-local mass and (angular) momentum in general relativity*. For a great many years, relativists had resigned themselves to the idea that the mass-energy of the gravitational field cannot be localized, and only the *total* energy of an asymptotically flat space-time can be assigned an unambiguous meaning. Twistor theory now allows us to do a good deal better (Penrose 1982, Penrose and Rindler 1986), though various difficulties remain. A full and up-to-date account is to be found in Paul Tod's article (1990—this volume; see also Mason and Frauendiener (1990), this volume), and it will not be necessary for me to go into the details here.

Nonetheless, one is compelled to confess that, so far, rather little that is both tangible and new has come through with regard to twistor theory's *original* physical aspirations. Let us now try to see how the theory stands with regard to each in turn of the above-mentioned questions.

1 The infinities of quantum field theory

Of the fundamental physical problems referred to in the opening paragraph, it is only the issue of infinities of quantum field theory that has been significantly addressed, so far. The main progress in this direction has been in the theory of *twistor diagrams* (see Hodges 1990 and Huggett 1990, this volume) which has been evolved as the twistor analogue of Feynman graphs. The intention has been that a procedure essentially equivalent to the conventional Feynman theory could be developed—except that it is intended that *finite* answers are to be obtained in important cases where the Feynman graphs diverge. The initial work in this area proceeded to a considerable extent by guesswork, analogy, geometrical considerations, aesthetics, and wishfull thinking (Penrose and MacCallum 1972, Penrose 1975b; cf. also Sparling 1975, Qadir 1978), but then later work (Hodges 1983a,b, 1985a,b 1990, this volume, cf. also Huggett 1990 this volume) not only put the twistor diagram theory on a sound basis, but also led to actual changes in which certain of the infinities of the standard theory have indeed become replaced by finite expressions.

As an initial step of the original scheme, the usual momentum states of the conventional theory are replaced by the (finite-normed) *elementary states* that arise naturally in twistor theory, so that the calculated amplitudes have some chance of being actually finite, rather than involving delta-functions, as is the case with momentum states. This allows, as a general objective of the twistor diagram formulation, that amplitudes might be computed over *compact* (high-dimensional) contours, the integrands being supposed to be analytic expressions at all points of the contours, so that the answers would accordingly be always guaranteed to be finite whenever these two requirements can be satisfied. However, since some of the answers, as 'correctly' computed by Feynman graph methods, are actually divergent, this entails that some change must be introduced into the procedures from those that would be obtained by direct translation of the corresponding Feynman graphs. This applies, as Andrew Hodges noted a good many years ago, even to some 'tree diagrams' of the standard Feynman theory, which are infra-red divergent. He was able to circumvent this problem in an ingenious way (Hodges 1985a), by replacing the previous factors $Z^\alpha W_\alpha$ that had ocurred in twistor diagram expressions according to

$$Z^\alpha W_\alpha \longmapsto Z^\alpha W_\alpha + k$$

where k is some (dimensionless) numerical constant whose value would be ultimately fixed by theory or experiment. At first, k merely provides a number whose logarithm enters into a finite expression which replaces each infra-red divergent quantity (the divergence being recovered when $k \to 0$), but k seems also to play a key role in eliminating ultra-violet divergences (Hodges 1985a) and it has a separate importance in relation to twistor diagrams for *massive* particles.

The factors $Z^\alpha W_\alpha + k$ bear some resemblance to factors

$$Z^\alpha X^\alpha I_{\alpha\beta} + m,$$

which Hodges uses in the twistor diagrams describing massive particles ('projection operators' for the mass eigenvalue m). These diagrams also make use of the so-called 'universal bracket factor' $[\ldots]_U$ (cf. Penrose 1979c, Hodges 1985b) which is 'defined' (formally) by the divergent expression

$$[x]_U = \ldots + (x)_{-2} + (x)_{-1} + (x)_0 + (x)_1 + (x)_2 + \ldots$$

where, for $n = 0, 1, 2, 3, \ldots$

$$(x)_{-n} = (x)^{-n}/n! \quad \text{(contour with boundary on x=0)}$$

and

$$(x)_{n+1} = -n!(-x)^{-n-1}/2\pi i \quad \text{(contour surrounding x=0)},$$

or by the formal expression, suggested by George Sparling (cf. Penrose 1979)

$$\oint [x]_U = \left\{ \oint \int_0^\infty + -\!\!\!\!\!\!\oint \ \oint \right\} \frac{e^z}{z+x} dz$$

Although the divergence difficulties are cleverly circumvented in Hodges's particular 'mass-projection' expression, a proper general understanding of the universal bracket is still lacking.

Many other problems in twistor diagram theory remain to be solved, but some good progress has been made towards the goal of finding a complete formulation of the *standard model* of weakly or strongly interacting particles in twistor diagram terms (Hodges 1990, this volume). One particular problem is to obtain a fuller understanding of the (high-dimensional) contours that occur in twistor diagrams. As noted above, these are supposed always to be compact (perhaps with boundary) so that the integrals will always be finite. The extent to which it has been possible to satisfy this compactness requirement so far has been definitely encouraging. However, particularly when mass is present, the status of this requirement is still unclear and it seems to demand the use of 'blown-up twistor space' according to which the line I, in projective twistor space **PT**, is replaced by a quadric surface. This is related to the 'googly twistor space' needed for the description of general relativity, and which will be described in outline below.

Another particularly important issue of twistor diagram theory is to understand, in purely twistorial terms, which twistor diagrams are to appear in any given process, and with what weighting (and sign) each diagram is to occur. A popular approach to the corresponding problem for Feynman diagrams, in modern quantum field theory, is to use the (formal) method of *path integrals*. However, an analogous procedure for twistor theory has not yet come to light. There is an apparently fundamental conflict between the twistor description of fields and that which is addressed by a path-integral approach. In the latter, paths are deliberately allowed in which the field equations are violated, whereas in twistor theory it is considered to be a virtue for the classical field equations to come out as solved *automatically* by the twistor descriptions! In twistor (diagram) theory 'off-shell' contributions in which field equations are violated come about in a different way (in effect, by the introduction of further twistors). The relation between this and path integrals has not yet come to light.

Another possible line of approach to the problem of 'twistor diagram generation' is through the ideas of a generalized conformal field theory involving 'pretzel twistor spaces' (Hodges, Singer and Penrose 1989, Penrose 1989), though this approach has not yet progressed very far. A 'pretzel twistor space' is a higher-dimensional analogue of a Riemann surface (with 'holes'), as occurs in standard conformal field theory (or string theory). Though perhaps superficially similar to a 'membrane' (or 'p-brane') theory (i.e. higher-dimensional string theory), this approach is fundamentally different in that the generalizations of Riemann surfaces are *complex manifolds*, either of three dimensions (projective case) or of four (non-projective), and in that the complex manifolds play the role of *twistor* spaces, in relation to space-time, rather than being regarded as being 'in' space-time. So long as general relativity is not involved, each such 'pretzel twistor space' would be a *flat twistor space* X, defined by the following properties, in the three-dimensional projective case:

1. X is a compact complex 3-manifold with boundary ∂X;
2. each component of ∂X is a copy of (i.e. is CR-equivalent to) the **PN** of standard twistor theory;
3. each point of $X - \partial X$ has a neighbourhood which is holomorphic to a neighbourhood of a line in $\mathbf{CP}^3$;
4. the canonical bundle of X admits a fourth root.

The reason for condition (4) is to enable the corresponding non-projective (four-dimensional) flat twistor space to be defined as the appropriate line bundle over the projective flat twistor space.

According to this proposal, a scattering process would be described by one of these flat twistor spaces (or by a linear superposition of processes described by different such spaces), where a positive or negative orientation would be assigned to each component of ∂X. Each positively oriented component would refer to an incoming particle state, and each negatively oriented component, to an outgoing particle state, where the in- and out-states (taken to be massless, in the first instance) would be described in the standard way by (1st) sheaf cohomology elements (restricted to **PN**). The procedure follows closely the one adopted in ordinary conformal field theory (cf. Segal 1990, Witten 1989). It is strongly motivated by the close analogy between the way (a) that the 'equator' S^1 (unit circle) divides the Riemann sphere into the 'northern hemisphere' S^+ and 'southern hemisphere' S^- and the way (b) that **PN** divides **PT** into $\mathbf{PT}^+$ and $\mathbf{PT}^-$. The splitting (a) of functions (i.e. H^0-elements) on S^1 into their positive and negative frequency parts according to whether they extend holomorphically into S^+ or S^-, is closely mirrored by the splitting (b) of solutions of the massless field equations into their positive and negative frequency parts according to whether the corresponding twistor functions (as H^1-elements) extend into $\mathbf{PT}^+$ or $\mathbf{PT}^-$. (This important fact realized one of the key original motivations behind twistor theory; cf. Penrose 1986a.)

The hope is that there should be some close relation between the construction of flat twistor spaces, their corresponding conformal field theories, and twistor diagrams. This would mirror the way that the early string theory showed how duality diagrams (Riemann surfaces with 'holes') could be used to make sense of the 'counting' of Feynman diagrams in strong interaction theory, and can also serve to replace certain collections of infinite Feynman diagrams by finite expressions. There does seem to be a corresponding role for twistor diagrams in relation to flat twistor spaces, but unfortunately this has not been explored very far as yet.

In conformal field theory, the in-states [out-states] are elements of Fermionic 'Fock spaces'

$$\mathcal{H} = \mathbf{C} \oplus H \oplus H \wedge H \oplus H \wedge H \wedge H \oplus H \wedge H \wedge H \wedge H \oplus \ldots$$

where each H is the space of positive-frequency [negative-frequency] functions (or sections of bundles) on a positively [negatively] oriented S^1 that constitutes a 'hole' boundary in the Riemann surface in question. In the case of a pretzel twistor space, the 'hole' boundaries are copies of $\mathbf{PN}$, and instead of functions, we have first cohomology elements, representing wave-functions of massless fields. We cannot interpret the elements of the higher-order spaces

$$H \wedge H, \quad H \wedge H \wedge H, \quad H \wedge H \wedge H \wedge H, \quad \ldots$$

as representing many-particle states, since each different particle taking part in a scattering process is to be represented by a different $\mathbf{PN}$ hole. Instead, the elements of these higher-order spaces (functions of several twistors) must presumably represent massive particles, in accordance with the twistor particle programme, that will be briefly described in the next section.

2 Symmetry and asymmetry in particle interactions

One of the most striking things about the twistor formulation—for good or for bad—is that by choosing the twistor space $\mathbf{PT}$ to be primary, rather than the dual space $\mathbf{PT}^*$ (or *vice versa*), we are led to an essentially *left-right asymmetric* description of physics. This would seem to be a desirable thing when we are trying to describe aspects of physics—notably weak interactions—for which such left-right asymmetry is known to be a fact of nature, but its desirability is more questionable for those interactions which are believed to be left-right *symmetric*. In particular, in the case of general relativity, we have a fundamental theory of space-time structure which is left-right symmetric, and this presents a severe challenge to any asymmetric twistorial description. It is a remarkable fact, however, that several new approaches to the description of standard general relativity have also been guided, for apparently quite independent reasons, into a left-right asymmetric formulation. These are approaches which relate, in one way or another, to what are known as 'Ashtekar variables' (Ashtekar 1988—see also Mason & Frauendiener 1990, this volume). The left-right asymmetry is expressed as an asymmetry between primed and unprimed 2-spinor indices—or, what amounts to the same thing, to an asymmetry between anti-self-dual and self-dual curvatures.

Twistor theory after 25 years – its physical status and prospects. *Twistors in Mathematics and Physics* Eds T. N. Bailey, R. J. Baston, 1–29. London Mathematical Society Lecture Note Series, 156, Cambridge University Press, 1990

The basic 'twistorial' reason for believing in a left-right asymmetric approach to physics (i.e. a preferance either for **PT** or for $\mathbf{PT}^*$ in the formulation) arises from the 'twistor-function' description of linear massless fields. A holomorphic function $f(Z^\alpha)$ (actually a representative 1-cocycle for an element of $H^1(\mathbf{PT}^+, \mathcal{O})$), which is homogeneous of degree $-n-2$, describes a wave function for a massless particle of helicity $n/2$ (n being an arbitrary integer). It is a striking fact that in this way we can automatically incorporate the two essential requirements for a massless one-particle wave-function, namely satisfaction of both the massless field equation and the positive-frequency condition. The fact that we are using a *holomorphic* function of the twistor Z^α (i.e. a 'function of Z^α' rather than a 'function of Z^α and $\overline{Z}_\alpha$') is the twistorial version of the basic quantum-mechanical requirement that ordinary wave functions must be functions just of position (or just of momentum) not functions of both position and momentum. For ordinary wave functions we can, if we prefer, choose functions of momentum instead of functions of position for our descriptions of quantum states, so long as we are consistent about this. Likewise, we can, if we prefer, consistently use holomorphic functions of *dual* twistors W_α (i.e. holomorphic functions of $\overline{Z}_\alpha$, where we simply relabel $\overline{Z}_\alpha$ as W_α, i.e. *anti*-holomorphic functions of Z^α. However, such an alternative choice must be consistent: it would make no sense to use, say, a position description for particles of positive electric charge and a momentum description for particles of negative electric charge; and likewise it would make no sense to use, say, a dual twistor description for massless particles of positive helicity and a twistor description of massless particles of negative helicity. (Such a description might have seemed tempting in view of the fact that right-handed—i.e. self-dual—non-linear gravitons seem to have a natural description in terms of dual twistors and left-handed—i.e. anti-self-dual—non-linear gravitons, a natural description in terms of twistors.) A particularly awkward aspect of any such attempt to describe massless particles in this hybrid way arises from the fact that there is often the need to describe massless particles which are not simply entirely right-handed or left-handed, such as plane-polarized photons. Thus, at least if we are describing massless particles, it seems to be necessary to make a choice in our twistorial representation: *either* a description in terms of twistors Z^α must be used *or* a description in terms of dual twistors W_α.

So long as we are concerned only with *linear* massless fields (without sources), this does not imply any serious left-right asymmetry for what it is possible to achieve with the twistor formalism, but, as is apparent with the the 'non-linear graviton construction' for (anti-)self-dual gravitational fields (Penrose 1976) and the Ward construction for (anti-)self-dual Yang-Mills fields (Ward 1977), the situation seems very awkwardly different for non-linear fields. (This raises the issue of the 'googly problem' which I shall return to later.) If it is supposed that Nature's ways actually accord with some of the basic ideas of twistor theory, and that she thus prefers, say, a twistorial description—or else she prefers a dual-twistorial description—then it would be expected that some left-right asymmetry should be present with the actual physics of non-linear massless fields. Of course, we already know that weak interactions are left-right asymmetric, but the above considerations should apply also to the gravitational field. The standard Weinberg-Salam-Glashow-Ward theory of unified electromagnetic and weak interactions implies that there is an indirect left-right asymmetry in electromagnetism, but the above considerations seem to imply a 'twistor expectation' of a left-right asymmetry in gravitation also.

Even for linear fields, there is twistorial left-right (or, rather, a self-dual/anti-self-dual) asymmetry in the case of fields with *sources*. For example, in the case of a Coulomb field, the twistor-function for the self-dual part would have the form

$$f(Z^\alpha) = \frac{1}{(Q_{\alpha\beta}Z^\alpha Z^\beta)^2}$$

while that for the anti-self-dual part would be something like

$$g(Z^\alpha) = \log(Q_{\alpha\beta}Z^\alpha Z^\beta)$$

or

$$g'(Z^\alpha) = \log\frac{Q_{\alpha\beta}Z^\alpha Z^\beta}{A_\gamma Z^\gamma B_\delta Z^\delta}.$$

In none of these cases do we get a global representation of the space-time field as an H^1 element in twistor space, but in the self-dual case we get a global description as a relative H^1 element (Bailey 1985). In the anti-self-dual case this does not seem to be so, however, and the situation is more obscure. Moreover, when we go over to the 'non-linear' Ward representation in terms of a line-bundle over in the anti-self-dual case we get 'charge quantization' and a non-Hausdorff bundle (Penrose and Sparling 1979, Bailey 1985). There is no analogue (as yet) in the self-dual case.

I have phrased the above discussion in terms of left-right asymmetry (i.e. parity P), since this is the most obvious of the discrete symmetry operations which convert left-handed massless particle into right-handed ones. However the operation C of charge-conjugation (particle-antiparticle interchange) also achieves this (witness the case of a neutrino), as do the operations CT and PT (where T stands for time-reversal symmetry). All of these four symmetries are violated in weak interactions, and it would appear that such violations could be well accommodated by twistor theory. In the context of the rules governing twistor diagrams, one only needs an asymmetry under interchange of black spots (twistors Z^α) with white spots (dual twistors W_α). But in addition, T and CP are known to be violated in K_0-decay, and this could arise, twistorially, out of some sort of asymmetry between $\mathbf{PT}^+$ and $\mathbf{PT}^-$. There are (controversial!) reasons for believing that even CPT should be violated in quantum gravity theory (cf. Penrose 1981, 1986b, 1989 and below). Such further symmetry violations would seem much more natural in the context of twistor theory than they do in standard space-time descriptions, but a good deal more understanding is needed if we are to see what the exact role of twistor theory in symmetry violation actually is.

Let us next consider massive particles, and how the quantum description of such particles is best to be incorporated into twistor theory. We recall that the twistor theory of massless fields is closely bound up with the quantized expressions for momentum and angular momentum for a massless particle:

$$p_a = \pi_{A'}\overline{\pi}_A, \quad M^{ab} = i\omega^{(A}\overline{\pi}^{B)}\epsilon^{A'B'} - i\overline{\omega}^{(A'}\pi^{B')}\epsilon^{AB}$$

where, in the twistor (as opposed to dual twistor) representation of massless wave-functions, we make the replacements

$$\overline{\pi}_A \longmapsto -\frac{\partial}{\partial\omega^A}, \quad \overline{\omega}^{A'} \longmapsto -\frac{\partial}{\partial\pi_{A'}},$$

in accordance with the standard twisor quantization rule

$$\overline{Z}_\alpha \longmapsto -\frac{\partial}{\partial Z^\alpha}.$$

(Here I am taking $\hbar = 1$.) These give the standard momentum and angular momentum operators when applied to a twistor function of the one twistor variable Z^α, the squared mass $m^2 = p_a p^a$ being identically zero. The procedure which is adopted in twistor particle theory in order to handle particles of *non*-zero mass is to replace the above expressions by sums

$$p_a = \sum_{i=1}^{r} \pi_{iA'}\overline{\pi}_{iA} \quad M^{ab} = \sum_{i=1}^{r}\{i\omega_i^{(A}\overline{\pi}_i^{B)}\epsilon^{A'B'} - i\overline{\omega}_i^{(A'}\pi_i^{B')}\epsilon^{AB}\}$$

where instead of acting on functions of just one twistor, these (quantized) operators now act on functions of several twistors

$$Z_1^\alpha, \ldots, Z_r^\alpha,$$

with

$$Z_i^\alpha = (\omega_i^A, \pi_{iA'}).$$

A twistor (wave-)function for a massive particle is now to be a holomorphic function of $Z_1^\alpha, \ldots, Z_r^\alpha$ (although the possibility that certain of these twistor variables might better be taken as dual twistors should not be overlooked). According to the original twistor-particle scheme (which I sometimes refer to as 'naïve twistor particle theory') *leptons* were to be the particles described by functions of just *two* twistors, say Y^α and Z^α, and *hadrons* by functions of *three* twistors, say X^α, Y^α and Z^α. One of the reasons for this was that the n-twistor internal symmetry group (the group of linear transformations of Z_i^α and $\overline{Z}_{i\alpha}$ that leaves p_a and M_{ab} invariant) is a slight inhomogeneous extension of $\mathrm{SU}(2)\times\mathrm{U}(1)/\mathbb{Z}_2$ (or $\mathrm{U}(2)$) in the case of $n = 2$, and a rather larger inhomogeneous extension of SU(3) in the case $n = 3$. These were basically the symmetry groups that arose in the standard classification of leptons and hadrons, respectively—in the 'good old days' before 'charm' was discovered! The 3-twistor scheme for the hadrons of those days provided quite a strikingly good fit, for the most part, although there were some notable anomalous multiplets, such as the nonet (rather than the expected octet) which involves the η^0, and certain families of resonances which seemed to have the wrong symmetries (see Hughston 1979). The 2-twistor scheme for leptons seemed to present more serious problems, since it provided for only two quantum numbers (in addition to spin and mass)—identifiable with charge and lepton number—leaving no way of distinguishing the muon from the electron. An ingenious suggestion due to George Sparling was that the required additional quantum number might be, in effect, the sign of the quantum number for total spin! He noted that the squared total quantum-mechanical spin J_2 of a massive particle, since it takes the form

$$J_2 = j(j+1)$$

(in units of $\hbar$), where j is the usual total spin quantum number, is invariant under

$$j \longmapsto -j-1.$$

Thus, for an observed total spin value, there are really two possible values of the quantum number j, namely j and $-j-1$. Sparling's suggestion was that we can allow for *negative* values of j , and the idea was that perhaps what distinguishes the muon from the electron was that one of these particles has $j = \frac{1}{2}$ and the other has $j = -\frac{3}{2}$. (There was some hint of support for this kind of idea in the expression for the spin operator in the 2-twistor scheme.) It may be that this suggestion has less plausibility now than it had at the time, owing to the further complication of the discovery of the τ-lepton. Nevertheless there does seem to be something in this idea, which shows up in the case of *hadrons.* It is (or was!) well known that if we plot the hadron resonances, for each particular set of values of its SU(3) quantum numbers, in a diagram with the j vertical and m^2 horizontal, then most of the resonances will lie on a family of remarkably straight lines rising off to the right—the *Regge trajectories.* In the case of the baryons, it is possible to divide these trajectories into two classes, those corresponding to natural parity ($\epsilon = 1$) and those corresponding to unnatural parity ($\epsilon = -1$). If we assume that (say) the natural parity baryons have positive j and the unnatural parity ones, negative j, and we now plot m^2 against j, we find that the pairs of Regge trajectories join together into single straight lines, the *extended* Regge trajectories (Penrose, Sparling and Tsou 1978)! There does seem to be some genuine physical content in this idea, since it has some predictive power and the later observations have been quite well in accordance with these predictions (Tsou 1981).

The ideas of (naïve?) twistor particle theory ought, in principle, to fit in well with the theory of twistor diagrams. For in order to know which interactions a specific particle is supposed to indulge in, i.e. which particular twistor diagrams are to be considered to contribute, there must be a way of specifying its appropriate *quantum numbers.* Twistor particle theory says that these quantum numbers ought to be expressible in terms of the eigenvalues of certain (scalar) twistor operators constructed from

$$X^\alpha, \ldots, Z^\alpha, \frac{\partial}{\partial X^\alpha}, \ldots, \frac{\partial}{\partial Z^\alpha},$$

as applied to a *twistor wave-function* $f(X^\alpha, ..., Z^\alpha)$, i.e. a function of a number of twistors, from which a particle's space-time wave function can be obtained by contour integration (see Penrose 1975a, Hughston 1979). The demand that f actually have a certain specified set of eigenvalues for these operators (i.e. that the particle whose twistor wave-function is f is supposed to be a particle of a specified type) can be given also in terms of certain types of twistor diagram. Here the diagram acts as a projection operator which projects out the part of f that has precisely the required eigenvalues.

It is now well understood that for twistor wave functions $f(Z^\alpha)$, of just one twistor variable Z^α, the function f is really to be thought of as a representative of *1st sheaf cohomology*, and the contour integration of f yields the evaluation of this cohomology element on lines in projective twistor space and provides a space-time wave-function for a particle of zero rest mass. When the number of twistors is greater than one, contour integration yields a wave-function for a particle which need not be massless, but it is not yet known, in the cases of twistor functions of more than one twistor, exactly how the function is to be interpreted in some cohomological sense. The most likely possibility seems to be in terms of relative cohomology, but more work needs to be done in order to decide what is really going on.

In the days before τ-leptons and charmed hadrons, etc. the twistor particle programme had seemed to fit in rather well with the observed particle classifications in a quite striking way. However things have got more complicated than they were, in the world of observational particle physics, and it would now seem that some essentially new ingredients are needed in order to make the twistor particle scheme work. If one had been strict about the original scheme, however, it would have been always clear that some new ingredient was needed. For the twistor mass operator commutes with everything else. This should imply that the mass of a particle, according to the scheme, could take absolutely any (positive) value, quite independently of the other quantum numbers. This is completely at variance with what is observed, since *observationally* the mass parameter seems to be fixed—or, at least, having only a very few descrete values allowed to it (e.g. those for the electron, the muon and τ-particle?). It would seem that the mass ought not to commute, exactly, with the other quantum numbers, but as yet, no plausible scheme for this has come to light. (Twistor particle theory is really not significantly worse than the standard schemes in this regard, there being no respectable theory at all for determing mass values for quarks, etc.)

It would seem that in order to understand how to treat mass properly, we must understand exactly how Einstein's gravity fits into the twistor scheme. Mass acts as the source for gravity, and it seems hardly likely that any deep understanding of mass can ever be obtained without general relativity being brought in in some essential way. (It would obviously be unreasonable to attempt to understand electric charge without bringing in Maxwell's theory!)

3 Coupling constants and particle masses

The problem of finding a theoretical determination of particles' mass values is part of the problem of understanding what fixes (pure-number) coupling constants in physical interactions generally. Although twistor theory has not yet provided any tangible advances towards an understanding of these deep questions, it is not impossible that it might ultimately do so. Perhaps one of the more plausible places where such an advance might be expected, assuming that there is some relevant significance to the (naïve) twistor particle programme, is in an understanding of the Cabbibo and the Weinberg angles. Both of these would have to do with the process of 'twistor growing' (Penrose 1976b). The Cabbibo angle would arise in the way that the 2-twistor scheme (for leptons) would be mapped into the 3-twistor scheme (for hadrons), and the Weinberg angle, in the way that the 1-twistor scheme (for the case of the photon) would be mapped into the 2-twistor scheme (for leptons). In each case, there would be a 'mixing' of twistors, in the scheme involving the larger number of twistors, when it gets reduced to the scheme involving the smaller number of twistors. The degree of mixing, in each case, would give rise to the 'angle' in question. 'Twistor growing' would be the reverse of this process, where the number of twistors gets increased in the passage from one scheme to the next. (This procedure would seem to have to do with something like a cycle: Čech cohomology → Dolbeault cohomology → 'complexification' by 'freeing up' the complex conjugates into separate twistors → hyperfunctional Čech description.)

The fine structure constant α, or equivalently the charge on the electron, is the most noteworthy example of a pure-number coupling constant. Particle mass-values, scaled in terms of the Planck mass, would be the gravitational analogues. In each case, the question of renormalisation comes in to muddy the issue. Should one be attempting to determine theoretically the 'bare' or 'dressed' values of these numbers? The bare value might seem more fundamental, but for this to be meaningful, we should need the renormalization factors to be finite. Thus we require a *finite*, and not a merely renormalizable theory—which is in accordance with the aspirations of twistor diagram theory (and of other fundamental approaches).

This notwithstanding, it has never been clear to me that it should be a 'hopeless' matter to attempt to determine the 'dressed'—or rather the 'observed' values—directly. It has often been argued, since all the processes of physics contribute to these renormalization factors, that there would be no chance of having an *exact* theoretical determination of them until one had a complete theory of physics—and, even then, the renormalization factors would have to be enormously complicated. However it not at all clear to me that this need be the case. There might be an alternative way of looking at quantum field theory according to which renormalization becomes completely irrelevant. In relation to all this, there is a certain 'numerology' in the observed pure-number values that arise, suggestive of possible laws operative on the scale of the dressed values themselves. Perhaps most noteworthy example of such numerology is Wyler's value of the fine structure constant, which I re-express as:

$$\alpha = (4\pi/3)^2(120\pi^3)^{1/4} = 137.0360824....$$

(the current observed value of α being 137.035895 ± 0.000061, which leaves Wyler's number only marginally outside the experimental range). If we take this numerology seriously, and think in terms of the 'elementary quark charge', namely one third of that of the electron, where we may take the factor of 4π as being 'natural' in the context of Gauss's law, etc., then we may regard the factor of $(4\pi/3)^2$ as being 'accounted for'—so what has to be explained is the $(120\pi^3)^{1/4}$. From the point of view of twistor theory, the unnatural-looking fourth root in this expression might somehow be related to the fourth root $Z^\alpha \rightarrow \{1 + f_{-4}(Z^\alpha)\}^{1/4} Z^\alpha$ that occurs in the 'googly photon' construction (Penrose 1979a), according to which the self-dual part of the Maxwell field can be coded in terms of deformations of the fibres in the map

$$\mathbf{T} - \{0\} \longrightarrow \mathbf{PT}$$

that preserve the volume form

$$d^4Z = \epsilon_{\alpha\beta\gamma\delta} dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta.$$

This is in contradistinction to the standard Ward construction for the anti-self-dual part, where the Euler operator

$$\Upsilon = Z^\alpha \frac{\partial}{\partial Z^\alpha}$$

is preserved instead. It would seem that in the correct twistorial description of interacting electromagnetic fields these right- and left-handed parts of the

Twistor theory after 25 years – its physical status and prospects. *Twistors in Mathematics and Physics* Eds T. N. Bailey, R. J. Baston, 1–29. London Mathematical Society Lecture Note Series, 156, Cambridge University Press, 1990

photon would have to be combined together, but it has been hard to see how to do this in such a way that the right-handed part does not come out as being 'charged' with respect to the left-handed part. In order to disentangle the two parts of the photon, it seems to be necessary to 'go to infinity' and thereby destroy conformal invariance in the description. As yet, an elegant way of formulating all this has not emerged.

The problem is intimately connected with the corresponding one for gravitation. As things stand, there is only a rather inelegant way of coding the self-dual part of the gravitational field into a deformation of twistor space, coding transformations like $Z^\alpha \to \{1 + f_{-6}(Z^\alpha)\}^{1/6} Z^\alpha$ into a 'blown-up' version of $\mathbb{T}$. Again, an elegant combination of such a 'googly graviton' (positive-helicity) with the standard 'leg-break' (negative-helicity) non-linear graviton construction (Penrose 1976a) would be needed (see later).

If such a combined construction were found, then it is conceivable that it might have a bearing on the problem of how the masses of particles are determined from their other quantum numbers. One should note that sourced fields (e.g. Coulomb, Schwarzschild) should, strictly, always involve both self-dual and anti-self-dual parts combined; and the quanta that are involved (photons, gravitons), being 'off shell', do not have well-defined helicities. Thus, there must be some more comprehensive treatment involving both 'leg-break' and 'googly' constructions in combined form, when sources are to be correctly incorporated. It is possible that global considerations in relation to this putative combined construction would lead to severe constraints on the strengths of the sources for these fields. Thus, it is conceivable that α, or basic particle masses, could be determined by the theory (and a corresponding comment would apply to the strong and weak coupling constants). Until such a combined construction becomes available, however, this must remain a matter of pure speculation.

4 Quantum gravity

Of course, quantum gravity was always very much in mind as one of the initial motivations for developing twistor theory. If something drastic happens to the structure of space-time at the very tiny scale of the Planck length, then some radical theory is needed in order to provide the necessary mathematical descriptions. Perhaps the very concept of a space-time point loses its meaning at such scales, in which case it is then necesary to find an alternative geometric description. This description would have to be consistent with the normal ones that we adopt on larger scales—and it could well provide better pictures, at those levels (concerning particles or atoms, or perhaps even molecules) where quantum effects become so important that they confuse our ordinary space-time view. But our description should also allow for the possibility that the very concept of a space-time point can evaporate in situations where quantum gravity effects become important, such as at the Planck length or at space-time singularities. The aspirations of twistor theory are very much in accordance with such goals, and the idea was mooted quite early on (Penrose 1975b) that 'at the quantum gravity level', instead of having a space-time manifold with a 'fuzzy *metric*' at the Planck scale (which had been a more conventional view), twistor theory would have 'fuzzy *points*', yet a well-defined concept of null direction. Some recent work (Mason 1990) has added some substance to this idea, but its proper implementation would have to wait for a more complete union between classical twistor theory and general relativity.

The conviction that classical Einstein theory must find an appropriate (and *elegant*) formulation within twistor theory received its main impetus from the *non-linear* ('leg-break') *graviton* construction (Penrose 1976a). Before that time, it had seemed that twistor theory was an essentially flat- (or conformally flat-) space theory, and that somehow gravitational effects were to enter in discrete 'quantized' elements which effected the 'glueing together' of pieces of Minkowski space (Penrose 1968). It had seemed that at the level of classical general relativity, twistor theory might lose its usefulness, and that gravitational effects might only be incorporated in terms of 'jumps' from one flat space to another. The central idea of the non-linear graviton—that twistor space could itself become 'curved' in a global way, the local structure being identical with that of flat twistor space, but that local curvature information for the space-time could be encoded in this global twistor structure—provided a new viewpoint and a new stimulus. There was now the hope that general relativity at the *classical* level could perhaps be completely transcribed into twistor terms. Indeed, it has become clearer, as time has gone on, that such a transcription is vital, if twistor theory is ever to move significantly further towards its goals, not only in providing a better foundation for quantum gravity, but in making significant headway towards resolving the other major obstacles to progress in physical theory, as described in the other sections of this account.

One argument indicating that the full (vacuum, at least) Einstein equations, rather than merely their linear approximation, needs to find an appropriate twistor description, if twistors are to have a chance of adequately describing quantized gravity, is that gravitational quanta ought not to be described merely by linear spin-2 massless states, but by (complex) solutions of the *full* vacuum equations. (See Penrose 1976; also, in the full Einstein theory, the gravitational radiation field at null infinity $\mathcal{I}$ involves non-linear complications in relation to the Bondi-Metzner-Sachs group, that do not occur with the linearized radiation field, cf. Penrose and Rindler 1986.)

At the time of writing, nearly fifteen years after the original non-linear graviton construction came to light, there is still no clear message as to how Einstein's great theory could be fully encompassed by a twistor formulation. There have been numerous ideas, but all remain rather incomplete as of now. Recent work on the 'ambitwistor' formulation of Einstein's equations has been encouraging (LeBrun 1990, this volume, Baston and Mason 1987). The points of (projective) ambitwistor space $\mathbf{P}\mathcal{A}$ represent complex null geodesics in complexified space-time, where $\mathbf{P}\mathcal{A}$ arises as a deformation of the (doubly) projective space of pairs (Z^α, W_α) for which $Z^\alpha W_\alpha = 0$. Here the Einstein equations are described in terms of the possibility of extending $\mathbf{P}\mathcal{A}$ into $\mathbf{PT} \times \mathbf{PT}^*$ to various infinitesimal orders. From the point of view of twistor wave functions, an ambitwistor description is not really what is wanted, since this would lead to picture more akin to a *classical* function on phase space (like $\Psi(x^a, p_a)$) rather than a quantum wave function (like $\psi(x^a)$ or $\tilde{\psi}(p_a)$). The non-linear graviton description is closer to what is required in this respect, since this indeed provides a 'non-linearization' of a twistor function $f_2(Z^\alpha)$, rather than of something like $F_{2,2}(Z^\alpha, W_\alpha)$ which would be the case for ambitwistors (suffixes indicating degrees of homogeneity). However, the problem as to how best to encode the 'googly' part of the non-linear graviton in terms of a non-linear version of a twistor function $f_{-6}(Z^\alpha)$ is not yet fully resolved.

The present status of the googly problem seems to be that there are indeed 'deformations' of twistor space that can encode the required information, but that it is not yet clear whether this is the appropriate mathematical procedure to adopt, since a natural construction for the corresponding self-dual complex space-time in terms of the googly-deformed twistor space has not yet presented itself. The nature of the proposed deformations is as follows.

I consider just the 'pure' googly problem, in which there is no anti-self-dual field or gravitational sources, and we have to try to encode the information of a self-dual gravitational field into a 'twistor space' $\mathcal{T}$ for which the projective twistor space $\mathbf{P}\mathcal{T}$ is flat, i.e. simply (part of) $\mathbf{PT}$, except that we need to blow up $\mathbf{PT}$, by replacing the line $\mathbf{I}$ (representing Minkowski infinity) by a quadric surface $\mathbf{PI}$, to obtain the blown-up (flat) projective twistor space $\mathbf{PT}_{\mathbf{I}}$. Here, the blown-up non-projective flat twistor space $\mathbf{T}_{\mathbf{I}}$ is defined as the space of

$$(\omega^A \pi_{B'}, \pi_{A'} \pi_{B'}),$$

where $\pi_{A'} \neq 0$, together with the limit points of that space. These limit points (given by $\omega^A \to \lambda^{-1}\omega^A, \pi_{A'} \to \lambda \pi_{A'}$ as $\lambda \to 0$) constitute the three-dimensional space $\mathbf{I}$, whose projective version is the quadric $\mathbf{PI}$. We now apply transformations $Z^\alpha \to \{1 + f_{-6}(Z^\alpha)\}^{1/6} Z^\alpha$ in a neighbourhood $\mathcal{K}$ of $\mathbf{I}$, giving transition function relations between overlaps of pairs of open sets of a covering the complement of $\mathbf{I}$ in $\mathcal{K}$. These transformations are the ones which locally preserve the map $\mathcal{T} \to \mathbf{P}\mathcal{T}$ and the form (regular on $\mathbf{I}$ because of the $\pi_{S'}\pi_{T'}$ factors):

$$\pi_{S'}\pi_{T'}\epsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta.$$

The idea is that the resulting space $\mathbf{T}_{\mathbf{I}}$ locally extends to $\mathbf{I}$, although it does not do so globally. (The construction provides something akin to an 'orbifold', but where there is *functional* freedom at the 'vertex' $\mathbf{I}$.) The cohomology class defined by f_{-6} apparently comes in the form of an element of relative second cohomology, and it is, in effect, 'exponentiated' to give the required local deformation of $\mathbf{T}_{\mathbf{I}}$. When used as an ordinary twistor function to deternine a massless spin-2 self-dual radiation field, f_{-6} would give the required self-dual gravitational radiation field at $\mathcal{I}^+$ as 'seen' from (future) timelike infinity i^+.

Of course, even if this programme were entirely successful, since it is based on the behaviour of the space-time at null infinity $\mathcal{I}$, it would still leave open the question of how best to describe *finite* space-time regions. We shall have to be able to describe such regions if, in particular, we are ever to obtain a twistorial understanding of space-time *singularities*—and thus come to terms with what is generally considered to be one of the major tasks of quantum gravity theory.

Indeed, one of the hopes for twistor theory is that it might be possible to provide a coherent description of the physics actually *at* a space-time singularity, where the normal physical concepts, being dependent on ordinary space-time ideas, become inappropriate. It is conceivable that a twistor space can still be used even when the concept of a space-time point has evaporated. One can imagine that some kind of construction analogous to that of the non-linear (leg-break) graviton whereby the space-time points arise as something like 'holomorphic curves' in a curved twistor space $\mathcal{T}$. Such 'curves' would then cease to exist where they would otherwise represent singular points, but the twistor space $\mathcal{T}$ could still be perfectly well defined. For such an idea to work it would certainly be necessary to find a description of (at least vacuum) solutions of the Einstein equations that need neither be self-dual nor anti-self-dual.

These remarks could apply to the 'generic' type of singularity that is expected to arise inside a black hole (or at the big crunch of a collapsing universe) where we anticipate that the Weyl curvature becomes infinite. However, the big bang, if we assume the *Weyl curvature hypothesis*, would have vanishing Weyl curvature, so we have conformal flatness and a simple flat twistor-space description there (cf. Penrose and Rindler 1986). The second law of thermodynamics is associated with a general tendency for the Weyl tensor to get larger with time. Thus, there is a tendency for the twistorial description of space-time to get more and more complicated as time unfolds!

If twistors are ever to provide a good description of a quantum gravity theory capable of coping with the physical problems provided by the existence of classical space-time singularities, then such a description must come to terms with this striking fact that the space-time singularities of our observed universe are indeed *time-asymmetric*. Accordingly, the theory of quantum gravity that twistor theory (or indeed any other theory, if successful agreement with our universe is to be obtained) must be time-asymmetric in its implications. According to my own views (cf. Penrose 1989), there would have to be an actual change in the rules of quantum theory as they are presently understood. The quantized description of gravity would not be a 'standard' one; it would gain its time-asymmetry from the fact that an objective *wave-function collapse* would need to be part of the the resulting theory. As I have stated elsewhere (Penrose 1986b, 1989) it is not implausible to suppose that non-linear effects will ultimately alter the very basis of quantum theory, leading to reduction of the state vector, as soon as the space-time geometry itself becomes subject to quantum superposition (cf. also Károlyházy 1974, Károlyházy, Frenkel and Lukács 1986, Komar 1969). However, mere non-linearity of the ordinary sort (such as the introduction of a non-linear term into Schrödinger's equation) would seem to be quite insufficient. For reasons indicated in the next section, I believe that a radical new viewpoint is needed with regard to the structure of space and time, and that twistor theory could supply an essential input towards finding such a viewpoint.

5 Quantum measurement

One of the most puzzling features of state-vector reduction is its non-local nature. As we know from experiments of the EPR (Einstein-Podolsky-Rosen) type, a measurement in one region of space-time can seemingly cause the state in a distant region to 'jump', the second space-time region being spacelike-separated from the first. The separation between the regions can be many meters, such as in the famous experiment of Aspect and his associates (cf. Aspect and Grangier 1986), or (presumably) even millions of light-years. If we are ever to obtain any kind of objective physical picture of what is going on in such situations, it would seem that any such picture would have to be constructed from essentially non-local ingredients. One of the early motivations for the introduction of a twistor-type formalism was, indeed, the hope that some non-local description of space-time might lead to a more comprehensible picture of reality in relation to puzzling non-local quantum phenomena like those of EPR.

Initially, twistor non-locality was evident simply because the space-time realization of a (projective) twistor is indeed a non-local object: an entire light ray, in the case of a null twistor, and something with also a spatial spread (a Robinson congruence) in the case of a non-null twistor. Conversely, a space-time point has a non-local description in projective twistor space: as a projective line. However, as our understanding of the correspondence between twistor space and space-time descriptions has grown, this non-locality has emerged in more subtle guises. It has been one of the most striking and characteristic features of the twistor representation of physical fields that local field quantities in space-time (Maxwell field strength, anti-self-dual Yang Mills curvature, free massless fields of general spin, the Dirac field, anti-self-dual space-time curvature, etc.) are represented in twistor terms as global structures (holomorphic bundles, 1st cohomology classes, global deformations, etc.) for which the *local* structure in twistor space remains identical with that in the field-free case. Here I refer to *holomorphic* local structure only and not to 'real' information such as the CR-structure of the imbedding of **PN** within **PT**. It is the hope that this pleasing and powerful feature will remain, for the twistor description of *all* physics at the basic level.

However, despite this fundamentally non-local character of such twistor descriptions, it is still not clear what light twistor theory sheds on the puzzling non-locality of EPR-type phenomena. In twistor theory, all basic physical phenomena seem to have a non-local twistor description. Even local space-time activities are described non-locally in twistor terms, so it becomes perhaps less surprising when something else that is twistorially non-local (such as a two-photon state in an Aspect-type EPR experiment) should turn out to have a *non*-local space-time description. In this connection I should perhaps mention the curious 'elemental states' described by Andrew Hodges which are, on the other hand, local in twistor space (H^0-elements) but their space-time translations are non-local (H^1- or H^2-elements). These are expressions like $(A_\alpha Z^\alpha)_n$ or $(B^\alpha W_\alpha)_n$ and they can be viewed as being essential ingredients of twistor diagrams. In space-time terms they provide kinds of 'cohomological wave-functions'.

All this perhaps begs the question of the real puzzle presented by quantum non-locality, which comes in only with the process of quantum *measurement*. From the point of view of twistor theory, there is another feature of quantum measurement—or of state-vector reduction R—apart from its non-locality, indeterminism, non-linearity, discontinuity and time-asymmetry, which distinguishes it from unitary evolution U, and this is the fact that it is *non-holomorphic*. As far as I can make out, all the processes of quantum mechanics that come under the heading of Ucan be expressed in holomorphic terms. Indeed, holomorphicity has proved to be an extremely useful tool, even in conventional space-time descriptions, particularly in relation to 'causality' and the positive frequency condition (dispersion relations, the analytic S-matrix; also the Riemann surfaces of string and conformal field theories). The rôle of holomorphicity is much more central in twistor descriptions, where global holomorphicity seems to take over even the very foundational rôle of differential equations in space-time physics. The fundamental ingredient of R, on the other hand, is the convertion of a probability amplitude λ into an actual probability according to the decidedly *non*-holomorphic operation

$$\lambda \to \lambda\bar{\lambda}.$$

The understanding of this holomorphicity/non-holomorphicity in twistorial terms must involve the introduction of new conceptions. I am inclined to believe that not only must gravity be involved in quantum measurement, but also that there is something odd about the way that *time* is involved in the Rprocess. Perhaps there is even some kind of a hint of a beginning of a space-time bifurcation going on, incorporating a partial reversal of time-directionality—which might go some way towards explaining the presence of complex conjugation here. Moreover, the very fact that we have probabilistic behaviour in the Rpart of standard quantum theory, and discontinuous 'quantum jumps', suggests that something radically new is needed in the translation from an essentially holomorphic twistorial description into a real space-time one. It seems to me to be not at all impossible that something like a 'quantum jump' might occur, not localizable in space-time terms, when distortions of twistor space might necessitate jumping from one family of objects in twistor space, representing space-time points, to another (say from one family of holomorphic curves in some curved projective twistor space $\mathcal{T}$ to a separate family of holomorphic curves in the same $\mathcal{T}$, according to the 'leg-break' non-linear graviton; cf. Penrose 1986b). All this would require the *full* Einstein theory to be translatable into twistor terms, if it is indeed true that the problem of quantum measurement, i.e. the process Ris indeed an essentially *quantum gravity* process. As I have stated above, I believe that a (non-linear) modification of the rules of standard quantum theory will be needed if Ris ever to be understood as physical process, and so that the classical world can emerge out of the quantum one. To date, twistor theory has been very conservative in this respect. All we have set ourselves up to do, in twistor diagram theory for example, has been to compute complex-number amplitudes for different physical processes. There has been no attempt to 'explain' why the procedure $\lambda \rightarrow \lambda\bar{\lambda}$ should give rise to probabilities, nor even to tell us under what physical circumstances this procedure should actually be applied. These are the problems of quantum measurement. Since, according to the above aspirations, this understanding should be part of the correct quantum gravity theory, it is indeed among twistor theory's ultimate tasks to address such issues also.

6 Where next?

How do I expect that twistor theory might develop in the future, if it is to make any real headway towards resolving any of these grandiose issues? It is, of course, risky to attempt predictions of this kind. For the developments that might be most clearly discernable at present are those, on the whole, that have already been most thoroughly explored. It is likely that the most important future developments will come through some new ideas that cannot now be adequately guessed. There are, however, some promising lines of development in addition to the ones that I alluded to in earlier sections, that have not in fact been explored very far as yet, mainly through lack of time.

A recurrent theme of the earlier sections has been the importance of finding an appropriate twistorial description of the full (at least vacuum) Einstein equations. The lack of such a description seems to be the major stumbling block to fundamental progress in almost all the areas that I have considered above. Moreover, the difficulty in finding one is, in my opinion, not due to an inherent inadequacy in twistor theory, but is indicative of the depth of the physical problems involved in finding the appropriate union between quantum theory and general relativity. Since the very conception of twistor theory was rooted in quantum–space–time ideas, it should not be surprising if it turns out that in order for twistors to catch up with classical general relativity, it will be necessary for them also to go beyond it.

There are some very recent developments concerning the description of space-times, particularly in relation to the Einstein equations, that have arisen through an attempt to bring hypersurface twistor theory and Ashtekar variables into closer relationship (Mason and Penrose 1989). Hypersurface twistors are defined with respect to an (analytic) hypersurface $\mathcal{H}$ (usually either spacelike or one of the null hypersurfaces $\mathcal{I}^-$ or $\mathcal{I}^+$, the spacelike case being considered here) in an (analytic) space-time $\mathcal{M}$. The points of the projective hypersurface twistor space $\mathbf{P}\mathcal{T}(\mathcal{H})$ are *α-curves* in the complexification $\mathbf{C}\mathcal{H}$ of $\mathcal{H}$, where an α-curve is a null curve with tangent vector of the form $p^a = \tilde{\pi}^A \pi^{A'}$ for which

$$p^a \nabla_a \pi_{B'} = 0.$$

In order that p^a be actually tangent to the α-curve, we must have

$$\tilde{\pi}^A = t^{AA'} \pi_{A'},$$

where t^a is normal to $\mathbf{C}\mathcal{H}$ in $\mathbf{C}\mathcal{M}$. The non-projective hypersurface twistor space $\mathcal{T}(\mathcal{H})$ is a line bundle over $\mathbf{P}\mathcal{T}(\mathcal{H})$ (usually without the zero section) where the fibres are given by the scalings for $\pi_{A'}$. Now, associated with any such $\mathbf{P}\mathcal{T}(\mathcal{H})$ (or $\mathcal{T}(\mathcal{H})$) is a complex anti-self-dual conformal 4-manifold $\mathcal{M}(\mathcal{H})$, referred to as *heaven-on-earth*, which is obtained from $\mathbf{P}\mathcal{T}(\mathcal{H})$ by the standard non-linear graviton ('leg-break') construction, the points of $\mathcal{M}(\mathcal{H})$ being holomorphic curves in $\mathbf{P}\mathcal{T}(\mathcal{H})$. A complex 3-submanifold of $\mathcal{M}(\mathcal{H})$, canonically identifiable with $\mathbf{C}\mathcal{H}$, consists of those points of $\mathcal{M}(\mathcal{H})$ which are described by the holomorphic curves in $\mathbf{P}\mathcal{T}(\mathcal{H})$ given by the families of α-curves passing through actual points of $\mathbf{C}\mathcal{H}$

The space $\mathbf{P}\mathcal{T}(\mathcal{H})$, or equivently $\mathcal{M}(\mathcal{H})$, provides a good part of the intrinsic and extrinsic geometry of $\mathcal{H}$, within $\mathcal{M}$, that would be necessary for the initial value problem for general relativity. However, it is lacking in two essential ingredients. One of these is the choice of *conformal scale* for $\mathcal{M}$, the construction of $\mathbf{P}\mathcal{T}(\mathcal{H})$ (and $\mathcal{M}(\mathcal{H})$) being entirely conformally invariant. The other is the actual location of $\mathbf{C}\mathcal{H}$ within $\mathcal{M}(\mathcal{H})$ (i.e. the specification of which holomorphic curves in $\mathbf{P}\mathcal{T}(\mathcal{H})$ actually represent points of $\mathbf{C}\mathcal{H}$). Lionel Mason and I have found that a convenient way of specifying both bits of information is in terms of a rank-two holomorphic affine bundle $\mathcal{B}$ over $\mathbf{P}\mathcal{T}(\mathcal{H})$. The fibre of $\mathcal{B}$ which lies above a point Z of $\mathbf{P}\mathcal{T}(\mathcal{H})$ is the space of solutions $\nu_{B'}$, given χ, of the equation

$$\pi^{A'} \nabla_{AA'} \nu_{B'} = \nabla_{AB'} \chi$$

on the α-plane in $\mathcal{M}(\mathcal{H})$ which meets $\mathbf{C}\mathcal{H}$ in the α-curve representing Z. (Note that $\mathcal{M}(\mathcal{H})$ contains α-planes since it is anti-self-dual, a separate α-plane passing through each α-curve in $\mathbf{C}\mathcal{H}$.) The ∇-operator and metric on $\mathcal{M}(\mathcal{H})$ are chosen to agree, at $\mathbf{C}\mathcal{H}$, with that on $\mathbf{C}\mathcal{M}$, and are such that the scalar curvature of $\mathcal{M}(\mathcal{H})$ vanishes. The scalar χ on $\mathcal{M}(\mathcal{H})$ is chosen so that

$$\chi = 0 \quad \text{on} \quad \mathbf{C}\mathcal{H}$$

where

$$\Box\chi = 0$$

in $\mathcal{M}(\mathcal{H})$, and where $\nabla^a\chi$ defines a *lapse function* on $\mathbf{C}\mathcal{H}$, the normal

$$t^a = \nabla^a\chi,$$

to $\mathbf{C}\mathcal{H}$ in $\mathcal{M}(\mathcal{H})$ being identifiable with that in $\mathbf{C}\mathcal{M}$.

The spinor $\nu^{A'}$ is taken to have homogeneity -1 in $\pi_{A'}$ (which allows the bundle $\mathcal{B}$ to be defined over $\mathbf{P}\mathcal{T}(\mathcal{H})$ rather than just over $\mathcal{T}(\mathcal{H})$ and χ is taken to have homogeneity 0. There is an additional structure to $\mathcal{B}$ defined by the bundle map $\mathcal{B} \rightarrow \mathbf{P}\mathcal{T}(\mathcal{H}) \times \mathbf{C}$, given by

$$\nu^{A'} \longrightarrow \pi_{A'}\nu^{A'} + \chi$$

(the expression on the right being constant on each α-plane of $\mathcal{M}(\mathcal{H})$), so there is a sub-bundle $\mathcal{B}_0$ defined as the kernel of this map (cf. Eastwood 1990). The holomorphic curves in $\mathbf{P}\mathcal{T}(\mathcal{H})$, over which lies a holomorphic cross-section of $\mathcal{B}_0$ define the points of $\mathbf{C}\mathcal{H}$. The necessary initial data at $\mathbf{C}\mathcal{H}$ for the evolution of Einstein's equations are then determined by the remaining structure of $\mathcal{B}$.

Thus far, the Einstein equations have played no special role in this construction, but it turns out that the constraint equations for the Einstein equations can be expressed quite concisely in terms of another bundle over $\mathcal{T}\mathcal{H}$ (or over $\mathbf{P}\mathcal{T}(\mathcal{H})$), closely related to $\mathcal{B}$, whose fibres consist of the local (dual) twistors U_α (of homogeneity -1) on the α-curves with local-twistor derivative along the α-curve, equal to

$$(0, \tilde{\pi}^A\nabla_A^{A'}\chi).$$

This bundle is an 'affinized' version of the cotangent bundle of $\mathcal{T}(\mathcal{H})$, and it is related to ambitwistor space $\mathcal{A}$. It is hoped that the evolution of the Einstein's equations should find a reasonably concise expression in terms of it. Further work is in progress.

What seems really to be required is a twistor space that applies globally to the description of an entire space-time, and which is freed from any dependence on a hypersurface like $\mathcal{H}$. The space $\mathcal{A}$ is not really the proper analogue of flat twistor space $\mathbf{T}$ (it has too many dimensions, and is really the anologue of the hypersurface $\mathbf{CN}$ in $\mathbf{T} \times \mathbf{T}^*$), and $\mathcal{T}(\mathcal{I}^-)$ or $\mathcal{T}(\mathcal{I}^+)$ though canonical, if $\mathcal{I}^-$ or $\mathcal{I}^+$, does not carry all the necessary information. Even if the 'googly' information is incorporated into $\mathcal{T}(\mathcal{I}^\pm)$, say according to the description indicated earlier (where sources, black holes, etc. are assumed to be absent), then we still have the difficulty of extracting the information about local space-time regions. This seems a formidable problem, and some way of encoding the Einstein evolution (perhaps incorporating the ideas indicated above) seems essential.

If any such twistorial description of general relativity turns out to be successful, then it is to be hoped that it would tie in with the expression for *quasi-local mass* (and angular momentum, etc.). See Penrose (1982), Penrose and Rindler (1986), Tod (1990 this volume). That expression is still in need of a better theoretical foundation, so that ambiguities in the definition and other matters may be cleared up. (See Mason and Frauendiener 1990, this volume, for some new ideas concerning its relation to the Hamiltonian for general relativity.)

One point concerning the further development of twistor theory, which is exemplified by the above ideas for a 'googly' construction, is that for many purposes it is necessary to consider the *non*-projective twistor space $\mathbf{T}$ or $\mathcal{T}$ as fundamental, rather than the projective twistor space $\mathbf{PT}$ or $\mathbf{P}\mathcal{T}$. This has already been a feature of the twistor particle programme for a long time, since the helicity operator, which defines homogeneity for a twistor function of just one twistor variable, is merely one among a list of different twistor operators defining particle quantum numbers. Moreover, in Hodges's introduction of the 'k' term in twistor diagram theory, we have another example of an essential rôle for non-projective twistor space. It seems to me to be likely that any adequate description of Einstein vacuum space-times will also be an essentially non-projective one. This would certainly appear to have to be the case if the optimistic hope of describing general solutions of the Einstein vacuum equations in terms of a single twistor space can ever be realized.

It is clear that in many respects twistor theory remains far from its goals; yet there are others in which it has achieved things that had not originally been anticipated, and there have certainly been enough positive signals along the route to suggest that the theory has a good deal more to contribute to fundamental physics. However, some of the most obstinate problems seem almost as intractable as ever. What is the best route to further progress? Perhaps more account should be taken of conventional successful theories than has been done so far, and some new and useful input can be thereby obtained. Indeed, one of the criticisms of twistor theory (and of twistor theorists) in the past has been that they have paid too little attention to the ideas that are currently popular or successful in particle physics and quantum field theory.

As a counter to such criticisms, I should draw attention to the fact that in the important developments of twistor diagram theory that have been made over recent years by Andrew Hodges, the standard models of electroweak theory and QCD have been the guiding influences, so that I think it can now be said that these models can essentially be completely described in twistor terms. It is true that path integrals, as such, have not yet had a great deal of influence on twistor theory, but there are certain fundamental difficultes in bringing these ideas together, as was indicated earlier. The currently popular ideas of conformal field theory have certainly had their impact on twistor thinking, as I have mentioned above, but there are still some severe difficulties in implementing these ideas.

Another line of thought that has been very popular in recent years, as part of the impact of modern quantum field theory on pure mathematics, is the idea of a *topological* quantum field theory. Here one is concerned with the quantized version of a 'field theory', such as $(2 + 1)$-gravity (cf. Witten 1988/9), in which there is no local field information and all the 'physics' is stored in global structure. This kind of situation is of interest *mathematically*, where one may be concerned with such questions as the global topology of manifolds or with knots or links, but its interest from the physical point of view would seem to be severely limited if the ambient space of the theory is to be interpreted as 'space-time' (as is normally intended to be the case). However, twistor theory sugggests a different tack. We are now familiar with the fact that local space-time field information is not stored locally in twistor space but globally. Thus, if we could re-interpret the ambient space of a topological quantum field theory as *twistor space* rather than space-time, then the above problem with physical interpretation would be turned to our advantage. Among the new difficulties that we now must face, however, is the fact that twistor space is *complex*, not real, so there is a whole host of new problems in connection with how one must interpret the 'complexification' of a real topological situation. A start has been made to the understanding of these issues in the concept of 'holomorphic linking', which seems to have some tantalising connections with twistor diagram theory (Penrose 1988. 1990b). There is even the hope that some conection might ultimately be made with an early, but almost forgotten, motivation behind twistor theory, that of *spin-networks* (Penrose 1971, 1972). The idea here was that there should be an underlying *combinatorial* foundation to physics, which has seemed to be at odds with the underlying *complex-holomorphic* foundation to twistor theory. However, these foundations may not be so far apart from one another as they seem. Combinatorial and holomorphic concepts can often enrich one another in important ways. In the idea of holomorphic linking there is a clear aspect of this, and it is conceivable that the long-awaited relationship between twistor diagrams and spin networks may come to light through this kind of idea.

Finally, the fact should should not be ignored that despite all the marvellous new insights that have come to us through quantum field theory, it is *general relativity* that is now *the most accurate theory known to science*! (Observations of the binary pulsar PSR 1913 + 16 have shown GR to be accurate to about one part in 10^{14} (T. Damour, private communication; see Weinberg and Taylor 1984) whereas the accuracy of QED, as exemplefied by the calculated accuracy of the magnetic moment of the electron, is known to only one part in 10^{11}. It is the challenge of coming to terms with Einstein's marvellous theory, in the form in which he actually gave it to us, that has engaged the attention of twistor theorists more than has any other aspect of physics.

References

Ashtekar, A (1988) *New perspectives in Canonical Gravity.* Bibliopolis, Naples.

Bailey, T.N. (1985) Twistors and fields with sources on worldlines. Proc. Roy. Soc. **A397**, 143-155

Baston, R.J. and Mason, L.J. (1987) Conformal gravity, the Einstein equations and spaces of complex null geodesics. Class. Quan. Grav. **4** 815-826.

Eastwood, M.G. (1990) The Penrose transform without cohomology, in Mason and Hughston (1990).

Ginsberg, M.L. (1983) Scattering theory and the geometry of multitwistor spaces. Trans. Amer. Math. Soc. **276**, 789-815.

Ginsberg, M.L. and Huggett, S.A. (1979) Sheaf cohomology and contour integrals. In Hughston and R.S. Ward (1979), pp.287-292.

Hodges, A.P. (1982) Twistor diagrams. Physica **114A**, 157-75.

Hodges, A.P. (1983a) Twistor diagrams and massless Moller scattering. Proc. Roy. Soc. Lond. **A385**, 207-228.

Hodges, A.P. (1983b) Twistor diagrams and massless Compton scattéring. Proc. Roy. Soc. Lond. **A386**, 185-210.

Hodges, A.P. (1985a) A twistor approach to the regularization of divergences. Proc. Roy. Soc. London **A397**, 341-74.

Hodges, A.P. (1985b) Mass eigenstates in twistor theory. Proc. Roy. Soc. London **A397**, 375-96.

Hodges, A.P., Penrose, R. and Singer M.A. (1989) A twistor conformal field theory for four space-time dimensions, Phys Lett. **B216**, 48-52.

Hodges, A.P. (1990) String amplitudes and twistor diagrams, an analogy. In Quillen, Segal and Tsou (1990).

Hodges, A.P. (1990b) this volume.

Hughston, L.P. (1979) Twistors and Particles. Lecture Notes in Physics No. 97 (Springer-Verlag, Berlin).

Hughston, L.P. and Ward R.S. (1979) Advances in Twistor Theory. (Pitman Notes in Mathematics, Pitman, San Francisco).

Isham, C.J., Penrose, R. and Sciama D.W., eds. (1975) Quantum Gravity, an Oxford Symposium, (Oxford University Press, Oxford).

Károlyházy, F. (1974) Gravitation and quantum mechanics of macroscopic bodies, Magyar Fizikai Polyoirat **12**, 24.

Károlyházy, F., Frenkel, A. and Lukács, B. (1986) On the possible role of gravity on the reduction of the wave function. In *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Komar, A.B. (1969) Int. J. Theor. Phys. 2, 257.

LeBrun, C.R. (1985) Ambitwistors and Einstein's equations Class. Quant. Grav. **2**, 555-63

Mason, L.J. (1990) Insights from twistor theory, in 'Conceptual problems in quantum gravity', ed. J. Stachel, Birkhauser.

Mason, L.J. and Frauendiener (1990) The Sparling three-form, Ashtekar variables and quasi-local mass (this volume).

Mason, L.J. and Hughston, L.P. (1990) Further Advances in Twistor Theory, Vol. 1: *The Penrose transform and ith applications.* (Pitman Notes in Mathematics 231, Longman, Harlow Essex, UK and John Wiley New York, USA.)

Penrose, R. (1968) Twistor Quantization and curved space-time Int. J. Theor. Phys. **1**, 61-99.

Penrose, R. (1971) Angular momentum: an approach to combinatorial space-time, in Quantum theory and Beyond, ed. Ted Bastin (Cambridge University Press, Cambridge).

Penrose, R. (1972) On the nature of quantum gravity. In *Magic without Magic*, ed. J.R.Klauder (Freeman, San Francisco).

Penrose, R. (1975a) Twistors and particles: an outline. In *Quantum Theory and the Structures of Time and Space*, eds. L. Castell, M. Drieschner and C.F. von Weizsäcker (Carl Hanser Verlag, Munich).

Penrose, R. (1975b) Twistor theory: its aims and achievements. In Isham, Penrose and Sciama 1975.

Penrose, R. (1976) Non-Linear gravitons and curved twistor theory Gen. Rel. Grav. **7**, 31-52.

Penrose, R. (1976b) How to grow (or at least half grow) new twistors from old. Twistor Newsletter **2**, reprinted in Mason and Hughston (1990).

Penrose, R. (1979) The backhanded (googly) photon, In Hughston and Ward (1979).

Penrose, R. (1979b) A googly graviton. In Hughston and Ward (1979).

Penrose, R. (1979c) The Universal Bracket Factor. In Hughston and Ward (1979).

Penrose, R. (1981) Time-asymmetry and quantum gravity. In *Quantum Gravity 2*, eds D.W.Sciama, R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Penrose, R. (1982) Quasi-local mass and angular momentum in general relativity. Proc. Roy. Soc London **A381**, 53-63.

Penrose, R. (1986a) On the origins of twistor theory Gravitation and Geometry *I. Robinson Festschrift*, ed. W. Rindler and A. Trautman, Bibliopolis, Naples.

Penrose, R. (1986b) Gravity and state-vector reduction. In *Quantum Concepts in Space and time eds. R.Penrose and C.J.Isham* (Oxford University Press, Oxford) 126-146.

Penrose, R. (1988) Topological QFT and twistors: holomorphic linking. Twistor Newsletter **27**; Holomorphic linking, postcript, *ibid.*

Penrose, R. (1989) *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, Oxford).

Penrose, R. (1990) Non-Hausdorff Riemann surfaces and complex dynamical systems, (in Mason and Hughston 1990).

Penrose, R. (1990b) Twistor particles, strings and links. In Quillen, Segal and Tsou (1990).

Penrose, R. and MacCallum, M.A.H. (1972) Twistor theory: an approach to the quantization of fields and space-time, Phys. Repts. **6C**, 241-315.

Penrose, R. and Rindler, W. (1986) *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry.* (Cambridge University Press, Cambridge).

Penrose, R. and Sparling, G.A.J. (1979) The anti-self-dual Coulomb field's non-Haussdorff twistor space, Twistor Newsletter **9**. In Mason and Hughston (1990).

Penrose, R., Sparling, G.A.J. and Tsou, S.T. (1978) Extended Regge Trajectories. J. Phys. **A11**, L231.

Perjes, Z. (1977) Perspectives of Penrose theory in particle physics. Rept. Math Phys. **12**, 193-211.

Perjes, Z. (1982) Introduction to twistor particle theory. In *Twistor Geometry and Non-Linear Systems*, ed. H.D. Doebner & T.D. Palev (Springer-Verlag, Berlin) Lecture Notes in Mathematics **970**,53-72.

Perjes, Z. (1983) Twistor internal symmetry groups Group theoretical methods in physics, ed. M.A. Markov, Nauka, English translation by Harwood Acad. Publ. Co. p.631-645.

Perjez, Z. and Sparling, G.A.J. (1979) The twistor structure of hadrons. In Hughston and Ward (1979) pp. 192-7.

Qadir, A. (1978) Penrose Graphs. Phys. Repts. **39C**, 131-67.

Quillen, D.G., Segal, G.B. and Tsou S.T (1990) The Interface of Mathematics and Particle Physics (I.M..A Conference Series, Clarendon Press, Oxford).

Segal, G.B. (1990) The definition of conformal field theory (*to appear*).

Singer, M.A. (1990) Twistors and four-dimensional quantum field theory. In Quillen, Segal and Tsou (1990).

Tsou S.T. (1981) Extended Regge trajectories updated. Twistor Newsletter **12**. In Mason and Hughston 1990.

Ward, R.S. (1977) On self-dual gauge fields, Phys. Lett. **61A**, 81-2.

Weisberg, J.M. and Taylor, J.H. (1984). Phys.Rev.Lett. **52**, 1348-50.

Witten, E. (1988) Quantum field theory, Grassmanians and algebraic curves. Commun.Math.Phys. **113**, 529-600.

Witten, E. (1988/89) 2+1 dimensional gravity as an exactly solvable system. Nucl.Phys. **B311**, 46-78.

198

TWISTORS, PARTICLES, STRINGS AND LINKS

The interface of mathematics and particle physics (1988), 49–58.
IMA Conf. Ser. New Ser., Vol. 24, Oxford University Press, New York, 1990

Various ideas are brought together here; see also Chapters 178, 182, and 183.

Twistors, particles, strings and links. *The interface of mathematics and particle physics* (1988), 49–58. *IMA Conf. Ser. New Ser.*, Vol. 24, 1990

3

Twistors, Particles, Strings and Links

R. Penrose

1. Basic twistor theory

Twistors provide a setting for physical theory that is different in several respects from the usual one. (See Huggett and Tod [7], Penrose and Rindler [14], [15], Ward and Wells [21] for further details.) In place of the conventional underlying geometry, namely a real four-dimensional space-time manifold, we have a complex manifold — whose complex dimensionality is either three (projective twistor space $\mathbb{PT}$, with the structure of $\mathbb{CP}^3$) or four (non-projective twistor space $\mathbb{T}$, with the vector space structure $\mathbb{C}^4$). The correspondence between space-time and twistor space is a non-local one, and questions of locality in space-time take on a new character when interpreted in terms of twistors. Instead of the normal description using space-time fields, which are essentially local objects (e.g. curvature, which can be of the Maxwell, Yang-Mills or Einstein variety), one has global information in twistor space: sheaf cohomology elements, in the case of the twistor description of linear fields, and global bundle structure or global manifold structure in the case of non-linear fields. A key ingredient of twistor theory is holomorphic (i.e. complex-analytic) structure. It is global holomorphic structure, in twistor theory, that replaces the local field equations of the normal space-time descriptions.

The most direct part of the twistor correspondence is the representation of entire light rays (null geodesics) in (conformally compactified) Minkowski 4-space $\mathbb{M}$ by single points in a five-real-dimensional submanifold $\mathbb{PN}$ of $\mathbb{PT}$. The equation defining $\mathbb{PN}$ is

$$\Sigma(Z) := Z^\alpha \bar{Z}_\alpha = 0, \tag{1}$$

where $\Sigma(Z)$ is a Hermitian-quadratic form (of signature $++--$) in terms of the twistor $Z \in \mathbb{T}$. (In the usual component description, $\bar{Z}_0$, $\bar{Z}_1$, $\bar{Z}_2$, $\bar{Z}_3$ are the complex conjugates of Z^2, Z^3, Z^0, Z^1 respectively.) The points of $\mathbb{PT}$ which do *not* lie on $\mathbb{PN}$ also have physical interpretations, namely as classical massless particles with varying energy and fixed (non-zero) spin, the *positive*-helicity particles being described by the points of $\mathbb{P}\mathbb{T}^+$ (given by $\Sigma(Z) > 0$ and the *negative*-helicity ones, by the points of $\mathbb{P}\mathbb{T}^-$ (given by $\Sigma(Z) < 0$). In fact, the points of $\mathbb{T}$ can themselves, up to a phase freedom

$$Z \mapsto e^{i\theta} Z, \tag{2}$$

be given a direct physical interpretation in terms of the momentum and angular momentum structure of classical massless particles (so the *modulus* of the scale factor for Z is thereby interpreted). The 4-momentum and angular momentum have explicit expressions, bilinear in Z^α and $\bar{Z}_\alpha$. The helicity of the particle then becomes just the expression $\Sigma(Z)/2$. The phase of Z, also, can be interpreted (up to a sign) in terms of a polarization direction for the particle.

2. A generalization of conformal field theory

When we use the twistor coordinates Z^α and $\bar{Z}_\alpha$ in place of the usual space-time positions and momenta x^a, p_a we find that space-time linear massless fields or, more appropriately, *massless single-particle wave functions*, have a concise description in terms of holomorphic sheaf cohomology of regions of $\mathbb{P}\mathbb{T}$(or of $\mathbb{T}$). A wave function for a massless particle of helicity $n/2$ is described in twistor terms by an element of the holomorphic sheaf cohomology group

$$H^1(\mathbb{P}\mathbb{T}^+, \mathcal{O}(-n-2)). \tag{3}$$

Here $\mathcal{O}(k)$ denotes the sheaf of holomorphic functions "twisted by k" (i.e. described by holomorphic functions of Z^α homogeneous of degree k). The fact that the "domain" of the cohomology element is $\mathbb{P}\mathbb{T}^+$ is what specifies the wave function as being of *positive frequency*. Elements of

$$H^1(\mathbb{P}\mathbb{T}^-, \mathcal{O}(-n-2)). \tag{4}$$

define complex fields ("wave functions") of *negative frequency*. We can also define corresponding cohomology elements on the CR-manifold $\mathbb{P}\mathbb{N}$, provided that we interpret this cohomology in the appropriate "CR-hyperfunctional" sense (cf. Andreotti and Hill [1], Penrose [11]), and we find, with this interpretation,

$$H^1(\mathbb{P}\mathbb{N}, \mathcal{O}(-n-2)) = H^1(\mathbb{P}\mathbb{T}^+, \mathcal{O}(-n-2)) \cup H^1(\mathbb{P}\mathbb{T}^-, \mathcal{O}(-n-2)). \tag{5}$$

This is closely analogous to the splitting of complex-valued functions — or, more appropriately, "half-forms", given by $\mathcal{O}(-1)$ — into their positive and negative frequency parts according to whether they extend into the northern or southern hemisphere:

$$H^0(S^1, \mathcal{O}(-1)) = H^0(\mathbb{C}^+, \mathcal{O}(-1)) \cup H^0(\mathbb{C}^-, \mathcal{O}(-1)), \tag{6}$$

CHAPTER 198

where S^1 denotes the equator ($= \mathbb{R} \cup \{\infty\}$) of the Riemann sphere, and $\mathbb{C}^+$ and $\mathbb{C}^-$ denote its (open) northern and southern hemispheres, respectively (and where for "equality" to hold in the above equation, rather than merely "inclusion", the first H^0 should be taken in the above hyperfunctional sense).

This analogy is suggestive of a way in which *conformal field theory* may be generalized to give a theory which could have direct relevance to interacting (massless) fields in four-dimensional space-time (Hodges, Penrose and Singer [6], Singer [19]). In place of the bounded Riemann surfaces of ordinary conformal field theory — i.e. generalizations of the region $\mathbb{C}^+ \cup S^1$ — we can consider three-complex-dimensional manifolds X which generalize the region $\mathbb{PT}^+ \cup \mathbb{PN}$. Any such manifold X, which we call a *flat twistor space* (or, more colloquially, a "pretzel twistor space"), is a complex three-manifold which is compact with boundary, each component of the boundary being a copy of $\mathbb{PN}$ (i.e. with the same CR-structure as $\mathbb{PN}$). We may take p of these $\mathbb{PN}$s to have positive orientation (representing in-states) and q of them to have negative orientation (representing out-states). The manifold X is to satisfy the requirement that each point in the interior of X has a neighbourhood which is biholomorphic to a neighbourhood of a $\mathbb{CP}^1$ in $\mathbb{CP}^3$ (and, as a technical side condition, that the canonical bundle of X admits a fourth root - so that the "non-projective" version of X can be constructed).

The now-standard procedures of conformal field theory (see Segal [18] and Witten [22]) show how, given any Riemann surface, compact with a number of parametrized boundary S^1s some positively oriented and some negatively, one may construct an interacting quantum field theory, where each "in-state" is an element of a Fock space associated with a positively oriented S^1 and each "out-state", an element of a Fock space associated with a negatively oriented S^1. The twistor generalization carries this construction over to the three-complex-dimensional case virtually without change.

What we end up with, given a flat twistor space X, is (up to an undetermined phase factor) a multilinear map from a product of Fock spaces to $\mathbb{C}$:

$$\mathcal{H} \times \cdots \mathcal{H} \times \mathcal{H}^* \times \cdots \times \mathcal{H}^* \longrightarrow \mathbb{C}, \tag{7}$$

where p is the number of $\mathcal{H}$'s in the above, and q is the number of $\mathcal{H}^*$'s. Here, $\mathcal{H}$ stands for the fermionic Fock space of free massless particles:

$$\mathcal{H} = \mathbb{C} \oplus H \oplus H \wedge H \oplus H \wedge H \wedge H \oplus H \wedge H \wedge H \wedge H \oplus \cdots, \tag{8}$$

where each element of H describes a single-particle state for a massless particle (or anti-particle). Here the helicity of the massless particle need not be specified, so all helicities are considered together. (On the other hand, we could choose to fix the spin to be $n/2$, but then we would have to replace $\mathcal{O}$ by $\mathcal{O}(-n-2)\oplus\mathcal{O}(n-2)$ in what follows.) Note that, as things stand, the Fock space is "fermionic" irrespective of whether the helicity is integral or half-odd integral. I shall have a comment to make about this in a moment.

The twistor description of a single-particle massless state is an element of $H^1(\mathbb{N},\mathcal{O})$ of positive frequency, i.e. an element of $H^1(\mathbb{T}^+,\mathcal{O})$. Thus we can take

$$H = H^1(\mathbb{T}^+,\mathcal{O}), \tag{9}$$

but this ignores the usual quantum-mechanical requirement that states be normalizable. In twistor terms, this normalizability condition is a somewhat awkward "L^2" condition on the cohomology elements. A stronger condition ensuring normalizability (but not completeness of the space of states), and which is sufficient for many purposes, is that the "domain" of the cohomology element extend to $\mathbb{T}^+ - \{0\}$, i.e. to that part of $\mathbb{T}$ which lies above the closure $\overline{\mathbb{PT}^+}$, of $\mathbb{PT}^+$. This is equivalent to requiring that the massless fields, in their space-time description, be real-analytic (including at infinity — so that, for example, momentum states are excluded). This would be the analogue, in ordinary conformal field theory, of a requirement that the functions on S^1 be taken to be real-analytic.

An essential feature of our generalization of conformal field theory to complex manifolds of higher dimension is that quanta are represented by H^1 cohomology elements, rather than just by ordinary functions (i.e. H^0 cohomology elements). In standard twistor theory, the fact that field quantities are represented by H^1s rather than by H^0s has important implications since H^1 elements can often represent infinitesimal deformations of complex structure. When such infinitesimal deformations are "exponentiated" to become finite deformations, it is sometimes possible to interpret the resulting deformed twistor spaces at providing the solutions of *non*-linear fields in space-time. The (anti-)self-dual Einstein and Yang-Mills fields have been successfully described in this way a number of years ago (Penrose [10], Ward [20]). It is to be expected that this feature will have an important role to play in the present theory also. (Some comments in relation to general relativity will be made shortly.)

3. Interpretation of Fock space elements

As with ordinary conformal field theory, there is a problem of how one is to interpret the various "many-particle" elements of the Fock space. What we are presented with is an amplitude (up to an overall phase factor) for each specification of p in-states and q out-states. We would

CHAPTER 198

like to think of the in-states as representing incoming particles and the out-states as outgoing particles, but since the in-states represent elements of an entire Fock space, and so do the out-states, we appear to have the possibility of a several-particle state for *each* individual in- or out-state, rather than just a one-particle state for each. In string theory, one takes the "one-particle" elements of the Fock space to represent a lowest mode of excitation for a single particle, and the "many-particle" elements to represent its higher modes of excitation, rather than representing several actual particles. It is proposed that we do something similar here, but now we are able to interpret these "higher modes" by taking advantage of the ideas of twistor particle theory (see Penrose [9], Perjés [16], [17], Hughston [8]).

In twistor particle theory, massless particles (of the normal type, where the Pauli-Lubański spin 4-vector is taken to be proportional to the momentum 4-vector) are described by holomorphic functions of a single twistor variable Z. Actually any such function — referred to as a "twistor function" $f(Z)$ — is a really just a representative function for a sheaf cohomology element, namely the element of $H = H^1(\mathbb{T}^+, \mathcal{O})$ that represents the particle's wave function, as referred to above. But loosely we can refer to $f(Z)$ as the twistor description of the massless particle's wave function. To get from the twistor function to the particle's space-time wave function ϕ we apply a contour integration, which we can write symbolically as

$$\phi = \oint f \tag{10}$$

(see Penrose and Rindler [15], Huggett and Tod [7] for details). How do we describe *massive* particles in twistor theory? Again we can use a twistor function, but now it must be a function of *several* twistor variables, e.g.

$$F(Y,Z) \quad \text{or} \quad G(X,Y,Z). \tag{11}$$

There are contour integral expressions for the space-time "field" obtained from the twistor function just as before, but (as of now) the correct cohomological interpretation of F, G etc. remains elusive. (The likelihood seems to be that F and G describe elements of *relative* sheaf cohomology, rather than of just ordinary sheaf cohomology.) For twistor functions of a single variable there is just one (scalar) quantum number, namely the *helicity*, and this is described twistorially in terms of the *homogeneity degree* of the twistor function (i.e. the helicity value $n/2$ arising from $\mathcal{O}(-n-2)$). For twistor functions of two variables, there are four scalar quantum numbers, two of which are the spin and mass. The remaining two can be given tentative interpretations in terms of electric charge and lepton number. For twistor functions of three variables, there are eight scalar quantum numbers, two of which are again spin and mass, and the remaining six of which can be tentatively interpreted in terms of hadronic $SU(3)$ quantum numbers. For more twistor variables we get still further quantum numbers.

The idea, then, is to interpret the "many-particle" elements of the Fock space as describing higher order "excitations of twistor space" which would correspond to single *massive* particles: leptons, $SU(3)$ hadrons, "exotic" hadrons, etc. Many puzzles remain to be sorted out concerning this. Not least among these is the fact that conformal invariance (for the space-time fields) is lost once we pass to the descriptions of massive particles. In twistor terms, this means that, for these higher excitations, the full symmetry of the space $\mathbb{PN}$ is lost, and a particular line I in $\mathbb{PN}$ representing infinity in Minkowski space $\mathbb{M}$, must be singled out.

4. Gravitational interactions

The question of how gravity and space-time curvature is to be represented in such a twistorial scheme also presents difficult problems. It is to be expected that the "flat" twistor spaces defined above will have to be generalized to become "curved" in some appropriate sense. The asymptotic twistor spaces of asymptotically flat space times (cf. Penrose and MacCallum [13]) are relevant models; but so far these provide appropriate descriptions of gravitons of one helicity only (the left-handed helicity, according to the usual twistorial conventions). There are also ways of incorporating the information of the opposite helicity into the structure of such a twistor space, but I do not think that these are yet fully satisfactory. (A fully appropriate description would require an adequate solution of what has become known as the "googly problem": the problem of incorporating the information of right-spinning gravitons in the structure of a twistor space, whose ordinary deformations encode only left-spinning gravitons.) Again, there must be a violation of conformal invariance in the deformations that are allowed.

It should be remarked that despite many grandiose claims that the "conventional" (super)string theories can provide us with what is referred to as a "consistent quantum gravity theory", such schemes do not at all come to terms with most of the profound issues that are involved in bringing together the principles of quantum theory and those of curved space-time geometry. Curved space-time is not really an integrated feature of the (super)string models that have been seriously proposed so far. All the detailed descriptions refer only to perturbation expansions about a flat space background, where many of the essential issues of principle are not even addressed. Moreover, the terms of the perturbation expansion involve only "linear gravitons" (spin-two massless quanta in flat space-time $\mathbb{M}$). There are good reasons for believing that such quanta do not give the correct descriptions for what would be the appropriate quantization of Einstein's general relativity. In Einstein's theory, gravitational radiation modes do not really decouple at infinity. Any burst of outgoing radiation will affect the very geometry of future null infinity, and so will influence the description of any burst of radiation that may be emitted later. This has serious implications for any theory of quantum gravity that is merely Poincaré invariant, and does not take into account behaviour under the BMS (Bondi-Metzner-Sachs) group (cf. Penrose and Rindler [15], Section 9.8). Related to this are certain infra-red problems, which are also intimately related to the BMS group (Ashtekar [2]; cf. also Penrose [10] for some different objections to the use of linear gravitons for a description of quantum gravity). The present proposal for a twistorial generalization of conformal field theory in principle allows genuine space-time curvature to be incorporated.

5. Twistor diagrams

Another important motivation for introducing this generalization is that it offers some hope for a systematic development of *twistor diagram theory* . This would be analogous to the way that string theories and conformal field theories provide a systematic and consistent way of describing scattering amplitudes where many Feynman diagrams would otherwise have to be summed up with ambiguities arising from "duality" equivalences. Twistor diagrams are the analogues, in twistor theory, of the Feynman diagrams of conventional theory. They have been systematically developed by Hodges ([3],[4],[5]), and can (at least in principle) provide an alternative to Feynman diagrams for computing amplitudes for scattering processes in quantum electrodynamics or standard electroweak theory. However, the procedures so far are rather more "ad hoc" than one would wish, and one has no entirely "twistorial" rule, as yet, for determining which twistor diagrams must be summed in a given scattering process. The hope is that our twistorial generalization of conformal field theory will eventually lead to such a rule, rather in the way that "duality diagrams" and the Riemann surfaces of string theory provide a consistent and systematic procedure for computing amplitudes.

6. Holomorphic linking

So far, not a great deal of progress has been made in finding a direct relationship between twistor diagrams and the building up of different kinds of pretzel twistor space (though it is possible to see certain basic connections). On the other hand, there is another set of ideas that have become popular amongst mathematical physicists, and which appear to have some significant relationship with twistor diagram theory. This is the theory of braids, knots and links, and there is some reason to expect that these ideas may ultimately tie in with the procedures for building up the different kinds of pretzel twistor spaces.

To understand why twistor theory might have relevance to knot theory, etc., consider the concept of a topological quantum field theory which, as Witten has shown, can provide fresh insights into the new knot and link invariants of Jones and others. The central feature of such field theories is the fact that there is *no local field information* (e.g. with the the case of "2+1-dimensional general relativity", as studied by Witten [22]), and everything of interest lies in the global structure. This is, of course, just what is wanted from the point of view of mathematical applications, these being of an essentially topological character. However, if we anticipate topological field theories having a direct relevance to *physics*, we are presented with a puzzle, since much physical information is normally described in terms of *local* field quantities, such as Einstein, Maxwell, or Yang-Mills curvature. This puzzle finds a natural resolution if we take the background space of the topological quantum field theory to be *twistor space* rather than space-time. As we have seen above, physical field information is coded *non-locally* in twistor space, and all local information actually disappears, as would be required in a topological theory. Accordingly, twistor space might indeed be more physically natural than space-time, as the background for a topological quantum field theory.

There is an immediate difficulty, however, in that twistor space is a *complex* space. If we try to think of what should be the complex analogue of a knot or a link in ordinary three-dimensional space, we are presented with a problem. A complex curve is *two*-dimensional, as a real manifold, and the ambient complex three-dimensional space would be *six*-dimensional, as a real manifold. Two-dimensional manifolds do not knot or link in six-space. However, with regard to linking, at least, there is the remarkable expression due to Gauss:

$$L(\mathcal{X},\mathcal{Y}) = (4\pi)^{-1} \oint_{\Gamma}\{((\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y}))\}^{-3/2}(\mathbf{x}-\mathbf{y})\cdot d\mathbf{x}\wedge d\mathbf{y},$$

where $\mathbf{x}$ and $\mathbf{y}$ are the position vectors (in Euclidean space $\mathbb{E}^3$) of independent variable points on two closed curves $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $L(\mathcal{X},\mathcal{Y})$ provides the linking number of the two curves. We can interpret this integral expression perfectly well when the curves $\mathcal{X}$ and

$\mathcal{Y}$ become complex and lie in an ambient complex Euclidean space $\mathbb{CE}^3$ and where the integration is over some two-dimensional (real) contour Γ in the product space of the complex curves $\mathcal{X}$ and $\mathcal{Y}$. Indeed, since we are concerned specifically with twistor space here, we shall regard $\mathbb{CE}^3$ to be embedded in a $\mathbb{CP}^3$ and we identify this $\mathbb{CP}^3$ with $\mathbb{PT}$.

The complex curves themselves need not be compact, but the real contour Γ is taken to be compact. Normally, Γ would be assumed to have topology $S^1 \times S^1$ with one S^1 on $\mathcal{X}$ and one S^1 on $\mathcal{Y}$, but the expression would also make sense with more general choices for Γ. The integral is well defined provided that Γ avoids the "zero-distance locus" $\mathcal{Z}$, given by the vanishing of the expression $(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})$ (although there is also a possible difficulty in making sure that the square root of this expression can be taken consistently over the whole of Γ). The value of $L(\mathcal{X},\mathcal{Y})$ depends upon the choice of an element of second homology, namely that represented by Γ, in the space

$$\mathcal{X}\times\mathcal{Y}-\mathcal{Z}.$$

However, the actual location of $\mathcal{Z}$ within $\mathcal{X}\times\mathcal{Y}$ is dependent upon a specific selection of (complexified) Euclidean metric on $\mathbb{CE}^3$. This has a certain arbitrariness from the point of view of $\mathbb{CP}^3$, involving selecting a particular "absolute conic" Ω lying in $\mathbb{CP}^3-\mathbb{CE}^3$. In fact the value of $L(\mathcal{X},\mathcal{Y})$ is not very sensitive to this arbitrariness, and any continuous variation of Ω which does not result in $\mathcal{Z}$ meeting Γ will leave it unchanged. The exact kind of "object" that is referred to by $L(\mathcal{X},\mathcal{Y})$ still remains somewhat mysterious, and more subtle issues are involved than just the "linking" that occurs in the real case.

These matters are currently being investigated (cf. Penrose [12]). Certain significant interrelationships with twistor diagram theory have emerged, but this is not the place to elaborate upon these developments. Various different-looking contour integral formulae arise, and it is to be hoped that (twistor-motivated) expressions of this kind may lead to some new insights concerning links and knots, quantum field theory, and the relationship between such matters and twistor theory.

References

[1] A. Andreotti and C.D. Hill, E.E Levi convexity and the Hans Lewy problem. Part I: Reduction to vanishing theorems; Part II Vanishing theorems, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **(3)26** (1972) 747–806.

[2] A. Ashtekar, Quantization of the Radiative Modes of the Gravitational Field, in *Quantum Gravity 2*, ed. C.J. Isham, R. Penrose and D.W. Sciama, Clarendon Press, Oxford, 1981, p. 416–438.

[3] A.P. Hodges, Twistor diagrams, *Physica* **114A** (1982) 157–75.

[4] A.P. Hodges, A twistor approach to the regularization of divergences, *Proc. Roy. Soc. London* **A397** (1985) 341–74; Mass eigenstates in twistor theory, *ibid*, 375–96.

[5] A.P. Hodges, 1989, in *this volume.*

[6] A.P. Hodges, R. Penrose and M.A. Singer, A twistor conformal field theory for four space-time dimensions, *Phys. Lett.* **B216** (1989) 48–52.

[7] S.A. Huggett, and K.P. Tod, *An Introduction to Twistor Theory*, London Math. Soc. Student Texts, L.M.S. Publ.,1985.

[8] L.P. Hughston, *Twistors and Particles*, Lecture Notes in Physics No. 97, Springer-Verlag, Berlin, 1979.

[9] R. Penrose, Twistors and particles: an outline, in *Quantum Theory and the Structures of Time and Space*, eds. L. Castell, M. Drieschner and C.F. von Weizsäcker, Carl Hanser Verlag, Munich, 1975.

[10] R. Penrose, Non-Linear gravitons and curved twistor theory, *Gen. Rel. Grav.* **7** (1976) 31–52.

[11] R. Penrose, Physical space-time and non-realizable CR-structures, *Proc. Symp. Pure Math*, **39** (1983) 401–22.

[12] R. Penrose, Topological QFT and Twistors: Holomorphic Linking, *Twistor Newsletter* **27**, 1988; Holomorphic Linking: Postcript, *ibid.*

[13] R. Penrose and M.A.H. MacCallum, Twistor theory: an approach to the quantization of fields and space-time, *Phys. Rept.* **6C** (1972) 241–315.

[14] R. Penrose and W. Rindler, *Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields*, Cambridge University Press, Cambridge, 1984.

[15] R. Penrose and W. Rindler, *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry*, Cambridge University Press, Cambridge, 1986.

[16] Z. Perjés, Perspectives of Penrose theory in particle physics *Rept. Math Phys.* **12** (1977) 193-211.

[17] Z. Perjés, Introduction to twistor particle theory, in *Twistor Geometry and Non-Linear Systems*, ed. H.D. Doebner and T.D. Palev, Springer-Verlag, Berlin, 1982, pp. 53-72.

[18] G.B. Segal, The definition of conformal field theory, 1989, *to appear.*

[19] M.A. Singer, in *this volume.*

[20] R.S. Ward, On self-dual gauge fields, *Phys. Lett.* **61A** (1977) 81-2.

[21] R.S. Ward and R.O. Wells, Jr., *Twistor Geometry*, Cambridge University Press, Cambridge, 1989, *to appear.*

[22] E. Witten, Quantum field theory, grassmannians and algebraic curves, *Commun. Math. Phys.* **113** (1988) 529–600; 2+1 dimensional gravity as an exactly solvable system, *Nucl. Phys.* **B311** (1988/89) 46–78.

MATHEMATICAL INSTITUTE, 24-29 ST. GILES', OXFORD OX1 3LB, ENGLAND.

199

MASS POSITIVITY FROM FOCUSSING AND THE STRUCTURE OF SPACE-LIKE INFINITY

Twistor Newsletter **31** (1990). With A. Ashtekar. Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. Chapman and Hall/CRC, London, 1995

The geometrical argument outlined in Chapter 193 is in need of some more careful examination of the behaviour of light rays near i^0, as is presented here.

102 Mass positivity from focussing and the structure of space-like infinity. *Twistor Newsletter* **31** (1990). With A. Ashtekar. Also published in *Further Advances in Twistor Theory,* Vol. 2. Eds L. J. Mason and L. Hughston. 1995

§II.3.11 Mass positivity from focussing and the structure of space-like infinity

by A.Ashtekar & R.Penrose (TN 31, October 1990)

In §II.3.10, one of us (RP) outlines an argument to show that in asymptotically simple space-times satisfying a certain assumption, the mass is non-negative. The assumption—called the ***null conjugate-point condition*** requires that every endless null geodesic contain a pair of conjugate points. This condition is not overly restrictive. For example, it is implied by completeness of null geodesics, a weak energy condition ($T_{ab}l^a l^b \geq 0$ if $l_a l^a = 0$) and genericity. The purpose of this contribution is to examine more closely what it is about the structure of spatial infinity, i^o, that the argument establishes. In particular, we will be able to establish that the mass in question is indeed the ***ADM mass at i^o***. More precisely, our main result is that the ADM 4-momentum P_a of an asymptotically simple space-time satisfying the null conjugate point condition is necessarily a (future directed) causal 4-vector at i^o. (In this article, the assumption of asymptotic simplicity will include, in addition to asymptotic flatness at null infinity, that at spatial infinity. Thus, we assume that the space-time is AEFANSI in the sense of Ashtekar (1980) and that the parity condition of Ashtekar (1985) is satisfied at i^o.)

Let us begin by reviewing the result presented in §II.3.10. Let $\mathcal{M}$ be an asymptotically simple space-time. Fix a point a^+ on the future null infinity, $\mathcal{I}^+$ on $\mathcal{M}$. Let it lie on the generator α of $\mathcal{I}$. The past-directed light rays from a^+ will be said to *focus negatively* if they meet $\mathcal{I}^-$ in a family of points which recede indefinitely into the past (i.e., towards i^-). Let us suppose that this occurs. Then, if we examine the light rays neighbouring a given ray λ through a^+, as λ approaches α, the total shear of these rays along λ has the form of a convergence in the radial direction and a divergence in the transverse direction (see figure 1.$\ominus$).

fig 1.⊖

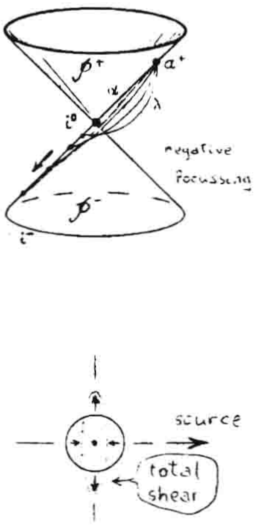
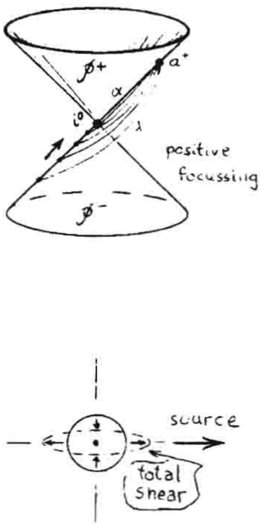


fig. 1.⊕



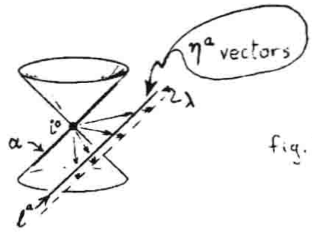


fig. 2a

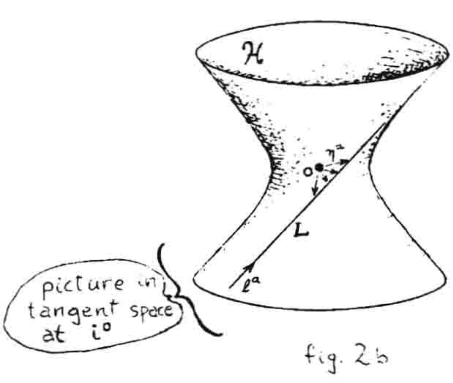


fig. 2b

Mass positivity from focussing and the structure of space-like infinity. *Twistor Newsletter* **31** (1990). With A. Ashtekar. Also published in *Further Advances in Twistor Theory*, Vol. 2. Eds L. J. Mason and L. Hughston. 1995

This is the behavior encountered in the negative mass Schwarzschild solution. In the positive mass Schwarzschild space-time, the situation is just the opposite. Now, we encounter ***positive focussing***. In this case, light rays from a^+ meet $\mathcal{I}^-$ in a family of points that approaches i^o, there is a divergence of rays in the radial direction and convergence in the transverse direction. Thus, if we move in radially, the rays spray away from each other as i^o is approached, whereas transversally they pinch towards one another. (See fig 1.⊕.) Now, as the null geodesic λ gets closer to α, various non-linear terms die off and the shear is given by the integral, $\int_\lambda ds\, C_{abcd} l^a l^c \eta^b \eta^d$, evaluated along λ, where l^a is the parallelly propagated tangent to λ, s is the affine parameter, and where η^a is the *radial* connecting vector. Thus, if negative focussing occurs, the integral would be positive and if positive focussing occurred, the integral would be negative. It was shown in §II.3.10 that, *if $\mathcal{M}$ satisfies the null conjugate point condition, negative focussing cannot occur.* Hence, in this case, the integral is necessarily non-positive for all geodesics λ originating at some point a^+ on $\mathcal{I}^+$ which are sufficiently close to the generator α of $\mathcal{I}$ on which a^+ lies. In this article, we show that this result in turn implies that the ADM 4-momentum is a (future pointing) causal vector. Thus, the intuition derived from the Schwarzschild solution is indeed valid more generally.

Let us recall the structure available at i^o. The asymptotic conditions of Ashtekar (1980) imply that the tangent space at i^o is well-defined and carries a (universal) Minkowskian metric of signature $(-+++)$. Further, along any C^1 curve with tangent vector η^a at i^o, $\Omega^{\frac{1}{2}} C_{abcd}$ admits a limit at i^o, where C_{abcd} is the Weyl tensor of the unphysical metric g_{ab}. We can decompose the limit into its electric and magnetic parts using the unit space-like vectors η^a at i^o, and thus acquire two *symmetric, traceless* tensor fields $E_{ab}(\eta)$ and $B_{ab}(\eta)$ on the hyperboloid $\mathcal{H}$ of unit space-like vectors at i^o. (Note that, by their definition, the two fields are tangential to $\mathcal{H}$). Let us focus on the asymptotic electric field, E_{ab}. It satisfies the field equation $D_{[a}E_{b]c} = 0$ on $\mathcal{H}$, where D is the derivative operator compatible with the natural metric $h_{ab} = g^o{}_{ab} - \eta_a \eta_b$ on $\mathcal{H}$, where $g^o{}_{ab}$ is the Minkowski metric in the tangent space at i^o. The information about the ADM 4-momentum P_a is contained entirely in E_{ab}:

$$P_a V^a := -\tfrac{1}{8\pi} \oint_S dS^a\, E_{ab} V^b, \tag{1}$$

where V^a is any vector in the tangent space at i^o and S is any 2-sphere cross-section of $\mathcal{H}$. The field equation and the trace-free property of E_{ab} imply that it is divergence-free while the projection of V^a into the hyperboloid (forced by its contraction with E_{ab} in the integrand) is a conformal Killing field on $\mathcal{H}$, whence the surface independence of (1). The field equation on E_{ab} also implies that it admits a scalar potential, E:

$$E_{ab} = D_a D_b E + E h_{ab} \tag{2.a}$$

In any given conformal completion, the potential E can in fact be constructed explicitly from the asymptotic (unphysical) Ricci tensor. The fact that E_{ab} is trace-free implies that E must satisfy the (tachyonic) massive scalar field equation:

$$D^a D_a E + 3E = 0 \tag{2.b}$$

on $\mathcal{H}$. In the Schwarzschild space-time with 4-momentum $P_a = mt_a$, with $t_a t^a = -1$, E is given by:

$$\frac{E = m\, 1 + 2(t.\eta)^2}{\sqrt{(1+t.\eta)^2}} \tag{3}$$

More generally, in physically interesting situations, the asymptotic magnetic field B_{ab} vanishes on $\mathcal{H}$ and the leading terms in the *physical* metric are dictated entirely by E. There exists a coordinate system in terms of which the physical metric has the asymptotic form:

$$d\hat{s}^2 = (1+\frac{E}{\hat{\rho}})^2 d\hat{\rho}^2 + \hat{\rho}^2\left(h^o{}_{\mathbf{ab}} + \frac{h^1{}_{\mathbf{ab}}}{\hat{\rho}} + \frac{h^2{}_{\mathbf{ab}}}{\hat{\rho}^2} + \ldots\right) d\hat{\phi}^{\mathbf{a}} d\hat{\phi}^{\mathbf{b}}, \tag{4}$$

with $h^1{}_{\mathbf{ab}} = Eh^o{}_{\mathbf{ab}}$. Here $h^o{}_{\mathbf{ab}}$ is the metric on the unit time-like hyperboloid in Minkowski space. Thus, $(\hat{\rho}, \hat{\phi}^{\mathbf{a}})$ should be thought of as the "asymptotic hyperboloid" coordinates.

It is easy to check that the space of solutions to the equation $D_a D_b \bar{E} + \bar{E} h_{ab} = 0$ is precisely 4-dimensional. The solutions are of the form $\bar{E} = K_a \eta^a$, where K_a is a fixed vector in the tangent space of i^o. Thus, there is some gauge freedom in the choice of the potential; we can add to the natural potential E of E_{ab} any $\bar{E}$ without changing the value of the field E_{ab}. This freedom is intertwined with the fact that if one uses only the asymptotic conditions of Ashtekar (1980), there is some ambiguity in the conformal completion at i^o. Given a completion, one can obtain a four parameter family of inequivalent ones by logarithmic translations. In the physical space language, there are the transformations of the type:

$$\hat{x}^{\mathbf{a}} \rightarrow \hat{x^{\mathbf{a}}} + K^{\mathbf{a}} log\hat{\rho}, \tag{5}$$

where $\hat{x}^{\mathbf{a}}$ are the asymptotically Cartesian coordinates, $\hat{\rho}^2 = x^{\mathbf{a}} x_{\mathbf{a}}$ and where $K_{\mathbf{a}}$ are constants. The new completions are C^1 related at i^o. Therefore, i^o and the tangent spaces in the two completions are naturally identified. Under this identification, the field E_{ab} of one completion is mapped to that of the second. The potentials E, however, are not preserved. On $\mathcal{H}$ we have:

$$E \rightarrow E + K_a \eta^a \qquad \text{and} \qquad E_{ab} \rightarrow E_{ab} \tag{6}$$

Since the field is unaffected, so is the ADM 4-momentum (and, to next order, also angular momentum). Thus, the logarithmic translations may be thought of as "gauge" in this framework. In a large class of space-times, this gauge freedom can be eliminated. Suppose, as in [3], that the electric field E_{ab} is reflection symmetric on $\mathcal{H}$. Then, we can demand that the potential should also be reflection symmetric. (Note that E of the Schwarzschild solution (eq 3) automatically satisfies the condition.

In Minkowski space the requirement singles out the potential $E = 0$.) This requirement selects a unique potential and hence removes the logarithmic ambiguity in the completion. As a part of the boundary conditions at i^o we assume that E_{ab} is reflection symmetric and work in a completion in which the potential E is also even under reflection.

With this formalism at hand, let us return to the implication of §II.3.10. As the null geodesic λ approaches the generator α in the completed space-time (fig 2.1), the connecting vectors η^a become, in the limit, the position vectors of points on a null geodesic (straight line) L in $\mathcal{H}$, and the tangent vector l^a is now parallelly propagated with respect to h_{ab} (fig 2.b). Since the argument of §II.3.10 tells us that integral of $C_{abcd}\eta^a\eta^c l^b l^d$ along λ is non-positive, in the limit we conclude that $\int_L ds\, E_{ab}l^a l^b \leq 0$. Using the expression (2) of E_{ab}, we have $E_{ab}l^a l^b = (l^a D_a)^2 E$, so that the last condition becomes:

$$\dot{E}^+ - \dot{E}^- \leq 0, \tag{7.a}$$

where $\dot{E} = l^a D_a E$ and ± denote, respectively, the values at the future and past (ideal) end points of L. Now, since E is reflection symmetric, the two terms on the left side of (6) add and we have:

$$\dot{E}^+ \leq 0. \tag{7.b}$$

To see the implication of this condition, let us examine the asymptotic form of E. Let us foliate the tangent space of i^o by a family of planes t=const and consider the corresponding foliation of $\mathcal{H}$. Assuming that E admits a power series expansion of the type $\sum E^{(n)}(\theta,\phi)\, t^{-n}$, where n runs from some *finite* negative value to $+\infty$, the field equation implies that E must admit an asymptotic expansion of the following type:

$$E(t,\theta,\phi) = (a_0 + a_m Y_{1m})(\theta,\phi))t + \frac{a_m Y_{1m}(\theta,\phi)}{2t} + \frac{E^{(3)}(\theta,\phi)}{t^3} + \ldots \tag{8}$$

where $Y_{1m}(\theta,\phi)$ are the three $\ell = 1$ spherical harmonics. (We believe the required assumption is always satisfied in the reflection symmetric case.) The condition $\dot{E}^+ \leq 0$ implies that the coefficient of the first term is non-negative and hence the 4-vector $\bar{P}_a = a_0 t^a + V^a$ at i^o–with t^a the unit time-like vector orthogonal to the slices and the spatial vector V^a in the $t = 0$ slice, given by $V^a\eta_a = a_m Y_{1m}(\theta,\phi)$—is future-directed and causal. At first, we were misled into thinking that the coefficient of the first term is the mass aspect at the future end of $\mathcal{H}$ and therefore the argument would show that the mass aspect should be positive. This is incorrect. In fact, the electric field E_{ab} constructed from the leading order term (via equation 2.a) vanishes identically whence the term makes no contribution what so ever to the ADM 4-momentum integral. Rather, the mass-aspect is the third term, $E^{(3)}$ in the expansion. However, because we have restricted ourselves to even potentials E, it *does* follow that the ADM 4-momentum P_a constructed from the correct mass aspect is precisely given by the vector $\bar{P}_a$, which resides in the leading order term. Therefore, although we cannot conclude that the mass aspect should be positive, (7.b) does indeed imply that the ADM 4-momentum is a causal, future-directed vector.

We conclude with two remarks. First, one can show that the vector space obtained by superposing the asymptotic mass aspects of Schwarzschild solutions (3) (whose the 4-momentum is not restricted to be time-like) is dense in the space of all asymptotic mass aspects $E^{(3)}(\theta, \phi)$ arising from smooth solutions to (2.b). We believe, furthermore, that the same is true of the entire solutions E everywhere on $\mathcal{H}$. Thus, in the reflection-symmetric case, one can work with superpositions of (3) without loss of generality. The second remark has to do with the leading term in the asymptotic expansion (8). The fact that this term diverges is an indication that there is a mismatch at i^o between the following two limits: sliding down a generator of $\mathcal{I}^+$, and, approaching i^o along a space-like direction and then making an infinite boost. The mis-match is a measure of the ADM 4-momentum. There is a similar mis-match in the way $\mathcal{I}^-$ is attached to i^o. This mis-match should show up in the metric coefficients rather than the curvature. If one allows E to acquire non-symmetric part under reflection, one can, by a logarithmic translation, remove a mis-match between i^o and $\mathcal{I}^+$. However, then the mis-match with $\mathcal{I}^-$ is twice as big. There is something "cohomological" here. Can twistor considerations shed further light on it? A clearer treatment of this issue is needed to make further progress in a proper understanding of asymptotically flat space-times, e.g. along the lines initiated by Friedrich.

We have presented here only the overall picture. Some details are yet to be worked out fully. Also, the results can probably be generalized in a number of ways. A more complete account will appear elsewhere.

References

Ashtekar, A. (1980) In: *General Relativity and Gravitation*, Vol. 2, (A.Held ed.), Plenum, New York, 1980.

Ashtekar, A. (1985) Found. Phys. **15**, 419.

200
TWISTOR THEORY FOR VACUUM SPACE-TIME: A NEW APPROACH

Twistor Newsletter **31** (1990). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

With this $\mathbb{TN}$ article, together with the subsequent one described in Chapter 203 (and earlier, in Chapter 202), a novel interpretation of how the notion of 'twistor' is to be understood physically, in terms of massless fields of helicity 3/2, this suggesting a direct link with the Einstein vacuum equations (Ricci-flatness—though there is no intended connection with the 'gravitino' of supergravity theory). Here, the idea is described in terms of Andrew Hodges's notion of an 'elemental state', which interprets a single twistor line connecting pairs of vertices of a twistor diagram. It may be noted that the cosmological constant Λ in Einstein's equations has to vanish here; later work by Paul Tod showed how to bring in a non-zero Λ.

108 Twistor theory for vacuum space-time: A new approach. *Twistor Newsletter* **31** (1990).
Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston,
P. Z. Kobak, and K. Pulverer. 2001

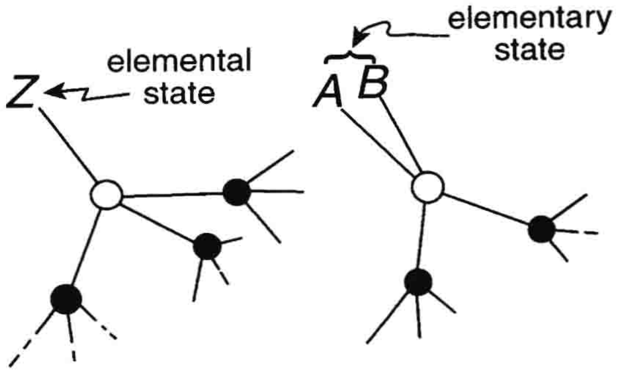
§III.4.19 **Twistor Theory for Vacuum Space-Time: A New Approach** *by R. Penrose* (TN 31, October 1990)

A fundamental problem that has confronted the twistor programme since its earliest days has been to find the appropriate twistor theory for curved space-times, and most particularly *vacuum* space-times. Many of the stumbling blocks to different parts of the general programme seem to be rooted in the absence of an adequate union between the ideas of general relativity and twistor theory. (See Penrose 1990 for a general survey of these problems and attempts at solution.) For some while it has seemed to me that one needs a concept of twistor with an intensely *non-local space-time description*. I shall present here a new concept of twistor that is, indeed, *very* non-local.

A difficulty with the asymptotic twistor/googly approach (with twistors defined at $\mathscr{I}$ only) has always been that twistors seem to have very little contact with the *finite* regions of space-time and with actual space-time points where, after all, the imposition of Einstein's vacuum equations $R_{ab} = 0$ is playing its role. The present suggestion may be regarded as a way of providing such contact.

The basic new ingredient is Andrew Hodge's concept of *elemental state*, §I.5.11, that, roughly speaking, is the 'one-eared' version of the normal 'two-eared' elementary state of standard twistor diagram theory.

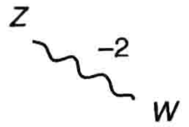
Twistor theory for vacuum space-time: A new approach. *Twistor Newsletter* **31** (1990).
Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston,
P. Z. Kobak, and K. Pulverer. 2001



The two ears of the elementary states represent the two singularities which give the twistor function its H^1 character. That is necessary in order that the $\oint$ can be performed so as to provide the corresponding space-time field $\varphi_{\dots}$. This field is an H^0, i.e., it has a *local* space-time interpretation. The appropriate '$\oint$' on a 'twistor function' with a single singularity (*one* ear) provides a *cohomological* space-time field. This is how the elemental states are described in space-time terms. There will be some covering (Stein) $\{\mathcal{U}_i\}$ of the relevant space-time region (say the complement of the α-plane represented by Z, in $\mathbb{CM}^\sharp$), and a field $\varphi_{\dots}$ in each $\mathcal{U}_i$, providing a representative function of the cohomological object in question.

The idea here is to try to represent a twistor Z^α in terms of the (dual) 'twistor function'

$$f(W_\alpha) = (Z^\alpha W_\alpha)_{-1} \qquad \text{(i.e.,}$$



)

or perhaps, making use of Andrew Hodge's insights with regard to twistor diagrams, something like

$$f(W_\alpha) = (Z^\alpha W_\alpha - k)_{-1}$$

Here $(x)_{-1}$ is the *bracket factor* satisfying

$$\frac{d^2}{dx^2}(x)_{-1} = \frac{1}{x}$$

so we can take $(x)_{-1} = x \log x - x$ (or $x(\log x + \gamma - 1)$, where γ is Euler's constant — which may be more in line with A.P.H.'s prescriptions), or else $(x)_{-1} = x$ together with some rule that the contour has boundary on $x = 0$ (or $x - k = 0$), or else use a limiting argument with complex powers. If we try to integrate out the W_α-dependence to get a space-time 'field', according to

$$\varphi_{\dots} = \oint (Z^\alpha W_\alpha - k(?))_{-1} d^2W$$

we end up with a *cohomological* field $\varphi_{\dots}$.

If f had been an ordinary (H^1-type) twistor function of homogeneity $n = 1$, then it would represent a helicity $s = \frac{3}{2}$ massless field ($n = 2s - 2$, for dual twistor functions), but since our f has the wrong cohomological structure for this, it represents a *cohomological* (or worse!) helicity $\frac{3}{2}$ field. The special feature of helicity $\frac{3}{2}$ for our purposes here is that the equations work (Buchdahl) also in Einstein vacuum space-times ($R_{ab} = 0$) — but only if we represent the field appropriately by a *potential,* which I shall take in the form

$$\boxed{\nabla_B^{B'} \gamma_{AB'C'} = 0, \qquad \gamma_{AB'C'} = \gamma_{AC'B'}}$$

where we take

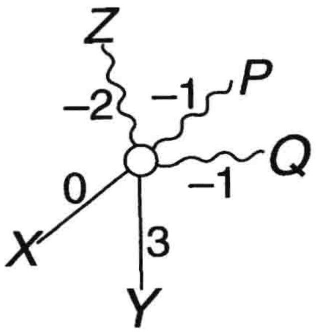
$$\boxed{\gamma_{AB'C'} \text{ modulo } \nabla_{AB'}\nu_{C'} \text{ with } \nabla_A^{B'} \nu_{B'} = 0\,.}$$

The consistency of these equations when $R_{ab} = 0$ is important to supergravity theorists, although the equations they use amount to the above without the symmetry condition on $\gamma_{\dots}$ and without the neutrino equation on ν. The two forms are locally equivalent (cf. Penrose and Rindler 1986, p. 370). An immediate (Buchdahl) consistency condition on the field equation for $\gamma_{\dots}$ is $R_{ab} = \lambda g_{ab}$, while we require $R_{ab} = 0$ in order that $\nu_{C'}$ act as a gauge.

The use of these spin $\frac{3}{2}$ equations here is totally different from their use in supergravity theory, however. The proposal here is to represent each twistor Z^α by an appropriate cohomology class (or 'worse'?) on a vacuum space-time $\mathcal{M}$, the representative functions being helicity $\frac{3}{2}$ fields as described above. Note that such a representation is *doubly non-local,* firstly because of the cohomology and secondly because the field must be represented as a potential modulo gauge. In flat space-time $\mathbb{M}$, we would have a local field quantity given by

$$\varphi_{A'B'C'} = \nabla_{A'}^{A} \gamma_{AB'C'}, \quad \text{satisfying } \nabla_A^{A'} \varphi_{A'B'C'} = 0$$

but in curved space-time $\mathcal{M}$, neither does the field equation for $\varphi_{\ldots}$ follow nor is $\varphi_{\ldots}$ independent of ν.



In $\mathbb{M}$, we can find such representative $\varphi_{\ldots}s$ and $\gamma_{\ldots}s$ by appealing to a twistor diagram. Here $X^\alpha = (ix^{AA'}\xi_{A'}, \xi_{A'})$ and $Y^\alpha = (ix^{AA'}\eta_{A'}, \eta_{A'})$ determine the field point (the line XY in $\mathbb{PT}$) and a local frame $\xi_{A'}$, $\eta_{A'}$ at that point; P^α and Q^α provide boundaries for the $\oint$ which determine the cohomological freedom. It appears to be the case that the field $\varphi_{\ldots}$ represents a *relative* H^2, which is taken relative to the α-plane determined by Z^α. The 'cohomological status' of $\gamma_{\ldots}$ is more obscure, because of the presence of logarithmic branching in its domain (work in progress!). In $\mathcal{M}$, where α-planes do not generally exist, in place of an α-plane, we seem to require a 'smeared out' or 'thickened' α-plane.

To see how this works explicitly in $\mathbb{M}$, take P and Q at infinity, i.e., of the form $P^\alpha = (\rho^A, 0), Q^\alpha = (\sigma^A, 0)$. Evaluating the above twistor diagram in $\mathbb{CM}$, we get a field (proportional to)

$$\varphi_{A'B'C'} = \frac{\pi_{A'}\pi_{B'}\pi_{C'}}{\sigma\cdot\omega(x)\rho\cdot\omega(x)} \qquad (\times\ \xi^{A'}\xi^{B'}\xi^{C'})$$

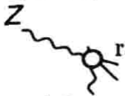
where ω^A and $\pi_{A'}$ are the spinor parts of Z^α at the origin 0 of $\mathbb{CM}$, with

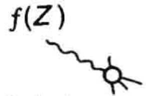
$$\omega^A(x) = \omega^A - ix^{AA'}\pi_{A'},$$

and $\sigma\cdot\omega(x) = \sigma_A\omega^A(x)$, etc. A potential for $\varphi_{A'B'C'}$ is given by

$$\gamma_{AB'C'} = \frac{\pi_{B'}\pi_{C'}\sigma_A \log(\rho\cdot\omega(x))}{\sigma\cdot\omega(x)}.$$

Of course, in $\mathcal{M}$, we cannot expect to find representatives of this form, and there is a problem as to how one should characterize those particular cohomological elements

that represent a *twistor* (elemental state), rather than something more general (i.e., 

rather than 

). The best I can do so far is to go to $\mathbb{C}\mathscr{I}^+$ (or $\mathbb{C}\mathscr{I}^-$) and demand that it relate to an α-line in the same way as the above expressions in $\mathbb{CM}$. However it would be better if we can avoid direct reference to $\mathbb{C}\mathscr{I}$, if possible. (Work in progress.)

Note that if we had a 'twistor' $Z^\bullet$ represented by $\gamma_{AB'C'}$ and a 'dual twistor' $W_\bullet$ represented by $\zeta_{A'BC}$, then we might be able to define their scalar product by appropriately integrating (over some Pocchammer-type contour) the 3-form

$$\Delta = \zeta_{F'}{}^{ED}\gamma_{E}{}^{F'D'}\varepsilon_{abcDD'}dx^a \wedge dx^b \wedge dx^c$$

(which is closed, and exact if either $\gamma_{\ldots}$ or $\zeta_{\ldots}$ is pure gauge). Andrew Hodges gives arguments to show that we should expect something like $\{\log(W \cdot Z/\hbar) + \gamma - 1\}W \cdot Z$ rather than just $W \cdot Z$. Some nonlinearities are certainly necessary. (Further work is in progress, e.g., relations to hypersurface twistors are being explored.) Thanks to many, particularly Andrew Hodges.

References

Penrose, R. (1990) Twistor theory after 25 years — its physical status and prospects, in *Twistors in Mathematics and Physics*, eds. T.N. Bailey and R. Baston, LMS lecture notes series, **156**, Cambridge University Press, Cambridge.

Penrose, R. and Rindler, W. (1984 & 1986) *Spinors and Space-Time, Vols. 1 & 2*, Cambridge University Press, Cambridge.

201

WHAT DOES THE BIG BANG TELL US ABOUT QUANTUM GRAVITY?

Mem. S A It. **62** (1991), no. 3, 607–14

WHAT DOES THE BIG BANG TELL US ABOUT QUANTUM GRAVITY?

Roger Penrose

University of Oxford, England

The Big Bang

There is now a good deal of astronomical evidence in favour of the standard theory of the origin of the universe, in which there was, some 10^{10} or so years ago, an extremely dense and very hot state into which the entire universe was initially concentrated and from which it has been expanding ever since: the big bang. Some of this evidence comes from observations of distant galaxies and quasars, which are receding from us as though from a gigantic explosion; some of it comes from a detailed examination of the actual nature of the material of which the universe is composed, which is consistent in minute detail with the presence of the initial hot and dense state; impressive additional evidence comes from observations of the background radiation, which pervades the whole of space at a black–body temperature of 2.7° absolute (2.7K), and is closely consistent with the detailed picture of the initial hot big bang–state.

There is also theoretical evidence coming from general relativity, as described by Einstein's equations. If we make the approximation that the universe is exactly spatially homogeneous and isotropic (this assumption itself being closely in accord with observations) then Einstein's equations can be solved exactly — as was first done by Friedmann in 1972 specifically for the case where the matter in the universe is approximated by a pressureless fluid. It turns out (taking the cosmological constant to be zero) that there is (at least one) space – time singularity in the metric, where the space–time curvature diverges to infinity. We may take this singularity to be in the past, and to represent the big bang. Thus, the equations of general relativity tell us that as we examine the space–time farther and farther into the past until the initial explosion is reached, we find that the curvature of space–time diverges to infinity. If the universe is taken merely to resemble a Friedmann (or similar) cosmological model in a rough qualitative way, the assumptions of exact homogeneity and isotropy being dropped, and also the assumptions governing the precise nature of the matter in the universe (such as it being a pressureless fluid) are removed also, save that the energy and pressure densities are to satisfy appropriate conditions of "physical reasonableness" then it still follows — now from the so–called "singularity theorems" (of Hawking and Penrose 1970) — that the universe, according to general relativity, must possess a singularity. Although one cannot infer directly from these theorems that the "singularity" is, like the big bang of the standard Friedmann models, a region of "infinite space–time curvature", it nevertheless still follows that the equations of classical physics have reached their limit — and, in a clear sense, general relativity predicts its own downfall as a complete description of the structure of space–time. If physics is to have any competence in explaining or describing what actually happened at the big bang, then this cannot be standard classical general-relativistic physics. It would seem that we need to incorporate the principles of quantum theory into our curved-space descriptions of physics.

The need for quantum gravity

It is indeed now generally accepted that it is quantum gravity that we shall require if we are ever to understand the precise nature of the big bang in its earlier stages: the first 10^{-43} seconds of the universe existence! Unfortunately, despite prodigious efforts on the part of a number leading physicists, we still appear to be a very long way from having any convincing theory of quantum gravity. The principles underlying Einstein's theory are in various ways incompatible with those of quantum theory. For example, even the fundamental quantum-theoretic relation between a particles momentum and the operation of infinitesimal translational symmetry becomes quite obscure when the space–time in fact possess no spatial symmetries. That particular problem is present even when one is merely concerned with applying quantum theory within a given curved background space–time and not with the quantization of the gravitational field itself. As soon as the gravitational field – as described by the space–time metric – ceases to be a classical structure, the very notion of causality (as determined by the locations of the null cones throughout space–time) must itself become quantized. Normally, classical space–time causality plays a fundamental role in any approach to relativistic quantum field theory (e.g. field operators are supposed to commute at spacelike but not at timelike separation). Even seemingly radical approaches to physics such as superstring theory do not at all come to terms with most of these issues of principle. I think that most physicists would accept that advances in quantum gravity theory to date have been problematic at best, and have achieved little towards providing any reliable new insight into the nature of the big bang.

This is certainly disappointing, but I do not feel that we should be discouraged. My own suggestion will be to turn the whole question around. Rather than trying to get an understanding of the big bang from some coherent theory of quantum gravity, let us try to use what we do know observationally about the structure of the big bang to provide us with insight as to the actual nature of quantum gravity. It seems to me that observations indeed tell us something quite unexpected about what the physical theory of quantum gravity must be like:

quantum gravity must be a time – asymmetric theory.

Evidence from the second law of thermodynamics

What observational fact about the big bang is there that can have such an implication? The thing I have in mind is not an observation of the big bang itself — directly referring only to an event that took place some 10^{10} years ago — but something that has profound relevance to physics now, namely the second law of thermodynamics. The second law has a certain "obviousness" about it when we think of applying it in the future temporal direction — for it says that systems will tend to "randomize" when left to themselves. It requires no special explanation that a given status should evolve in the future into something more random – this being the increase of entropy that the second law asserts, entropy being a measure of randomness. However, the law presents a distinct puzzle when we think in its action in the reverse direction of time. In the past direction things would seem magically to proceed to successively more and more organized (less random) states, with entropy reducing more and more, as time progresses farther and farther into the past. To explain this, we need the early universe to be highly constrained in some very special non–random (low entropy) way. If we can accept that for some reason there was a state of very high organization given to us in the initial state, then the second law need present us with no additional puzzle. This law arises because, at the beginning of time, the universe was in an extremely special state (extremely low entropy) and has been continually randomizing (entropy increasing) ever since.

What was the nature of the "extreme specialness" at the big bang? At first sight there seems to be a paradox in the fact that in the standard model of the big bang the universe is taken to be in a state of thermal equilibrium, which means maximum entropy! Indeed, the very detailed agreement that the standard model has with the observations of nuclear abundances and with the black–body nature of the background radiation depends upon this assumption of thermal equilibrium in the initial stages of the universe. It is often claimed that the paradox is removed by the fact that the universe was very tiny in these initial stages and, accordingly, there was a low ceiling on possible entropy values that the universe might initially have. On this view, as the universe expanded, the ceiling on the entropy correspondingly increased, and the actual entropy of the universe began to lag behind this maximum permitted entropy — the second law arising because the actual entropy value is always "striving" to catch up with this permitted maximum.

However, this argument is not correct: for it fails to take into account the entropy potentially available in the degrees of freedom of the gravitational field, i.e. in the *irregularities of space – time geometry* — an entropy that can completely swamp any other form of entropy. (If it were correct, it would imply that the second law would have to be violated in the collapsing phase of a recollapsing universe.) These gravitational degrees of freedom describe deviations from uniformity in the spatial structure of the universe — and it is *these* degrees of freedom that were, for some reason, set to zero in the big bang. Thus the very precise specialness of the big bang was in fact that it was extremely *uniform*. The initial state was "thermal equilibrium" *only* with respect to the degrees of freedom in the *matter* in the universe. The gravitational degrees of freedom were *not* thermalized and could remain unthermalized (i.e. the universe could remain fairly uniform) for a long time owing to the extreme weakness of the gravitational coupling. Only gradually, as matter began to clump into galaxies and stars, did the gravitational field begin to move away from this special configuration and to take up some of the entropy available to it.

In fact it is precisely this gravitational clumping that gives us the entropy imbalance — or "lowness of entropy" — that is made use of by life on earth. For the sun is a hot spot in an otherwise dark sky, this hot spot having been created by the fact that the sun has been able to concentrate itself into a small hot object from a previously uniform distribution of diffuse gas. Green plants, and ultimately animals including ourselves, make use of this very imbalance in order to lower their own entropy, via the process of photosynthesis.

Space–time curvature near singularities

How can we understand the special nature of the big bang singularity in geometrical terms? Let me first make some remarks about space–time curvature generally. The full curvature is called the *Riemann* curvature, and it has twenty independent components. It can be split naturally into two parts of ten independent components each: *Ricci* curvature and *Weyl* curvature. The physical meaning of the Ricci part can be described in terms of the initial relative (accelerative) change in the *volume* of a small sphere that is released from rest and allowed to fall freely under gravity. The meaning of the Weyl part is (more or less) that it measures the initial *tidal distortion* of such a sphere. Einsteins' equations assert that the Ricci part of the curvature is directly determined by the mass-energy distribution of the matter. The Weyl part describes the free gravitational field, and it is more indirectly related to the matter distribution — rather in a way that the Maxwell field tensor is determined by the charge–current vector of (charged) matter.

How do we characterize the space–time geometry near a singularity in terms of these two types of curvature? Observations (particularly those of the 2.7K background radiation)

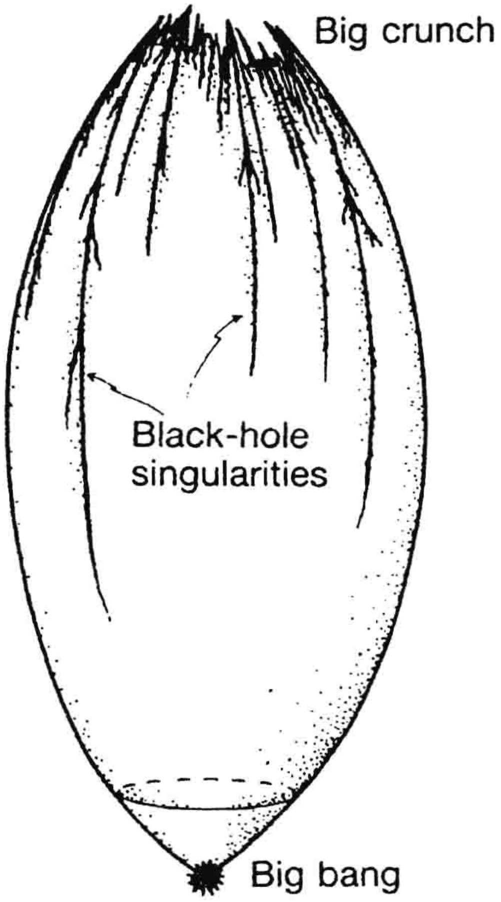

Big crunch
Black-hole singularities
Big bang

Fig. 1 Space–time picture of a "realistic" closed universe subject to the second law of thermodynamics — satisfying the Weyl curvature hypothesis.

indicate that the early universe was very close to being homogeneous and isotropic. Had it been exactly so, the Weyl part of the curvature would have been exactly zero. In fact, it appears that the closer we approach the big bang, the closer to being zero the Weyl tensor was. Observations are consistent with an assumption that the Weyl tensor was indeed exactly zero at the big bang itself, while the Ricci curvature diverged to infinity. We may take this as the initial constraint that characterizes the extreme specialness — or low entropy — of the big bang. It is the vanishing of the initial Weyl curvature that describes the fact that the gravitational degrees of freedom were set to zero initially.

On the other hand, when matter collapses into a black hole a space–time singularity is finally encountered in which the Weyl curvature would be expected to diverge to infinity (and to dominate over the Ricci curvature). This also would be the expectation for the high–entropy "generic" singularity that would be encountered in the final big crunch singularity of a closed universe — which would be the congealing of many black holes. (See Figure 1 for a description.) The hypothesis that any initial space–time singularity must be constrained by the Weyl tensor vanishing whereas final singularities are not so constrained, I call the Weyl curvature hypothesis. This hypothesis is consistent with observations, and would explain why we have a second law of thermodynamics in the form in which it has been given to us. The Weyl curvature hypothesis is grossly time–asymmetric. It — or something very like it — would have to be an implication of our sought–for theory of quantum gravity.

How special was the big bang?

It is worth pointing out just how enormous the constraint on the big bang needs to have been, so that something like the Weyl curvature hypothesis appears to be essential. For argument's sake, let us take a universe containing 10^{80} baryons. (The universe is at least this big; but for larger numbers of baryons, the figures would be even more striking than those that we shall be coming to in a moment.) By using the Bekenstein – Hawking formula for the entropy of a black hole — according to which the entropy of a black hole is, when measured in natural quantum–gravitational units, exactly one–quarter of the surface area of its horizon — we can estimate the total entropy of this universe as it finally collapses to its big crunch. The figure is 10^{123}, which is enormously in excess of the entropy — about 10^{88} that would be present in the 2.7K black body radiation at the universe's early stages and which had previously seemed to represent a puzzling huge figure.

Owing to the logarithmic relationship between entropy and phase–space volume, we can now estimate the "specialness" of the big bang as

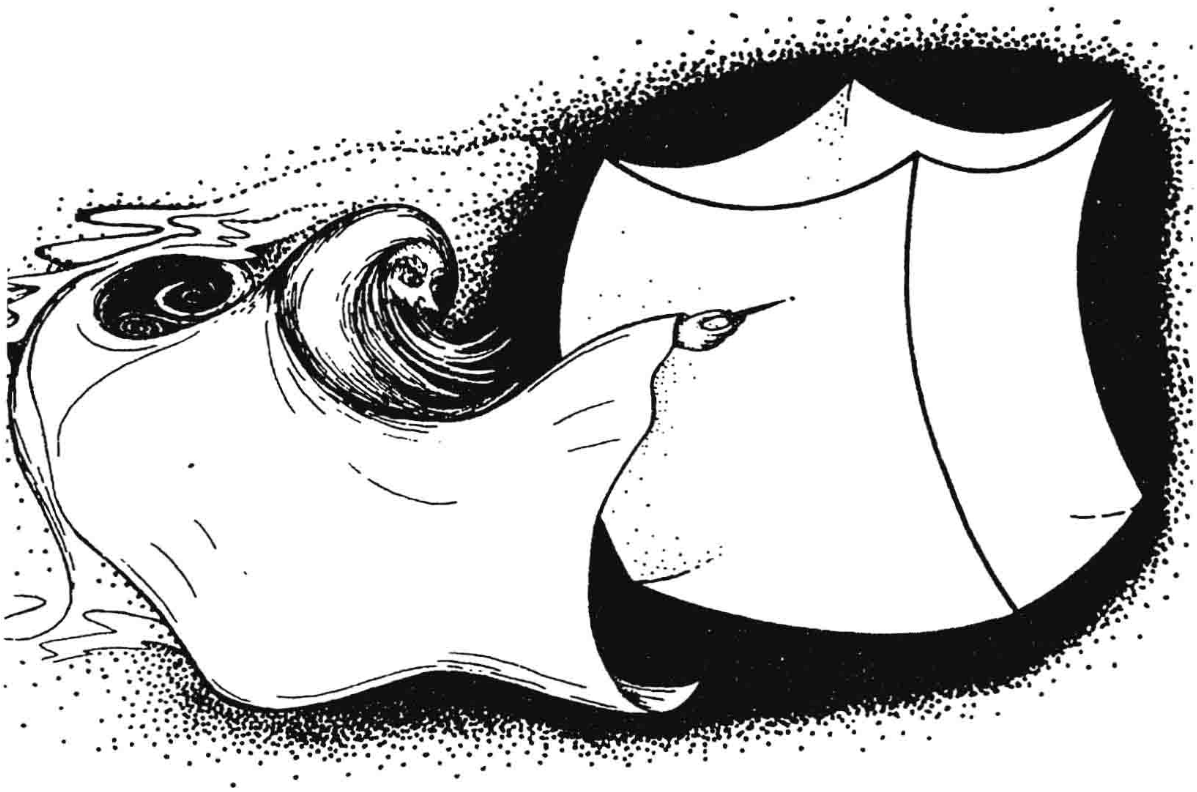


Fig. 2 The Creator locating a universe (with 10^{80} baryons) in accordance with the second law, with an accuracy of one part in $10^{10^{123}}$ (not to scale!).

$$\underline{\text{one part in }} 10^{10^{123}}.$$

This is the accuracy that would be needed for the "Creator" to locate this universe, subject to the Weyl curvature hypothesis, within the phase space of all possibilities (see Figure 2). According to my own "scientific' point of view, this is what quantum gravity must do for us!

State–vector reduction as a quantum gravity effect?

How can we expect to find such a theory of quantum gravity so grossly time–asymmetric that the Weyl curvature hypothesis could be one of its consequences? General relativity is time–symmetric, and so is the Schrödinger equation. However, one can give persuasive arguments that the process of state – vector reduction (or collapse of the wave–function) is a grossly time–asymmetric phenomenon. Many physicists appear to believe that he state–vector reduction is an "illusion" of some sort, but in my own opinion it must be a real physical effect. One can also provide an argument suggesting that there is a relationship between the time–asymmetry of state–vector reduction and the time–asymmetry of the Weyl curvature hypothesis: accordingly, each one of these phenomena would be "the other side of the coin" from the other. If this is so, then reduction would tend to come about as soon as gravity (i.e. changes in space–time structure) begins to become significant in a quantum process. The relevant scale would be that of the Planck mass, which is about 10^{-5} grams. Whenever enough of the environment is disturbed that masses of this order are involved, then the reduction process could set in.

As yet there is no detailed theory that makes coherent sense of all these ideas. Such a theory would have to involve non – local ingredients of a kind that would present severe difficulties of interpretation in normal space–time terms. (See Penrose 1989 for much further discussion.) The creation of such a theory presents a fundamental and profound challenge for physics

References

Hawking, S.W. and Penrose, R. (1970): "The singularities of gravitational collapse and cosmology", *Proc. Roy. Soc. (London)*, **A314**, 529–548.

Penrose, R. (1989) "The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics" (Oxford Univ. Press, Oxford)

202

TWISTORS AS SPIN 3/2 CHARGES

Gravitation and modern cosmology: The cosmological constant problem (1990), 129–37, *Ettore Majorana Internat. Sci. Ser. Phys. Sci., 56*, Plenum, New York, 1991

Here (and in Chapter 203) the twistor concept of Chapter 200 is re-interpreted as a *charge* for the massless field of helicity 3/2.

124 Twistors as spin 3/2 charges. *Gravitation and modern cosmology: The cosmological constant problem* (1990), 129–37, *Ettore Majorana Internat. Sci. Ser. Phys. Sci.*, 56, 1991

Twistors as Spin 3/2 Charges

Roger Penrose

Mathematical Institute
Oxford, UK

Abstract It is pointed out that twistors play a role as the charges for helicity 3/2 massless fields. Since such fields can be defined consistently in general Ricci-flat 4-manifolds, a possible new approach to defining twistors in vacuum space-times is indicated.

Introduction

In this note I point out a curious and apparently significant new role for twistors, as the *charges* for massless fields of helicity 3/2. This gives another angle on a proposal that I have recently introduced (Penrose 1990) aimed at finding an appropriate twistor concept for general Ricci-flat space-times. (That approach was partly stimulated by certain ideas due to Hodges 1990a,b.)

It is perhaps appropriate, for two reasons, that I honour Peter Bergmann's 75th birthday in this way. In the first place, it was Peter's paper with Ray Sachs (Sachs and Bergmann 1958) that made clear how momentum and angular momentum play an essential role in Minkowski space-time as the corresponding charges for a massless field of spin 2, the linearized gravitational field. Secondly, this analogous role for twistors, in the case of spin 3/2, offers considerable hope that the construction that I put forward in 1975 in honour of Peter's 60th birthday (Penrose 1976), which has become known as the "non-linear graviton construction" for the (anti-)self-dual solutions of the (complex) Einstein vacuum equations, might by the use of these ideas find an appropriate generalization in which the (anti-)self-dual restriction would be removed. (See Penrose 1980 for an outline of the twistor concepts needed.)

With regard to the latter considerations, this account must be regarded as exploratory and provisional. There remain certain profound and puzzling issues yet to be resolved. Those are matters for the future. I shall therefore concentrate here largely on the situation in flat (Minkowski) space $\mathbb{M}$, though I shall also indicate what is involved in the generalization to a curved Ricci-flat space-time $\mathcal{M}$.

1 Helicity 3/2 massless fields in $\mathbb{M}$ and their charges

I use the formulation given in Penrose (1969) (see also Fierz 1940) for (classical) massless fields in $\mathbb{M}$, according to which the gauge-invariant) *field* is a valence-3 spinor $\psi_{A'B'C'}$ subject to

$$\psi_{A'B'C'} = \psi_{(A'B'C')},\ \nabla^{AA'}\psi_{A'B'C'} = 0. \tag{1}$$

Conventions and notation are as in Penrose and Rindler (1984, 1986); the use of primed indices here rather than unprimed ones ensures that, for a positive-frequency wavefunction, $\psi_{A'B'C'}$ describes spin 3/2 particles of positive rather than negative helicity, i.e. *helicity* 3/2.

We can, locally, find a potential $\gamma^{C}_{A'B'}$ for $\psi_{A'B'C'}$ subject to

$$\gamma^{C}_{A'B'} = \gamma^{C}_{(A'B')},\ \nabla^{AA'}\gamma^{C}_{A'B'} = 0 \tag{2}$$

where

$$\psi_{A'B'C'} = \nabla_{CC'}\gamma^{C}_{A'B'}, \tag{3}$$

the gauge freedom being

$$\gamma^{C}_{A'B'} \longmapsto \gamma^{C}_{A'B'} + \nabla^{C}_{B'}\nu_{A'} \tag{4}$$

where $\nu_{A'}$ satisfies the Weyl anti-neutrino equation (helicity $\frac{1}{2}$)

$$\nabla^{AA'}\nu_{A'} = 0. \tag{5}$$

We also have, locally, a "second" potential $\rho^{BC}_{A'}$ subject to

$$\rho^{BC}_{A'} = \rho^{(BC)}_{A'},\ \nabla^{AA'}\rho^{BC}_{A'} = 0 \tag{6}$$

where

$$\gamma^{C}_{A'B'} = \nabla_{BB'}\rho^{BC}_{A'} \tag{7}$$

with gauge freedom given by

$$\rho^{BC}_{A'} \longmapsto \rho^{BC}_{A'} + \nabla^{C}_{A'}\chi^{B} \tag{8}$$

when χ^{B} satisfies the Weyl neutrino equation (helicity $-\frac{1}{2}$)

$$\nabla_{BB'}\chi^{B} = 0. \tag{9}$$

Let us suppose that there is a field ψ (indices suppressed), subject to (1), in some region $\mathcal{R}$ of $\mathcal{M}$, surrounding a world-tube that we suppose contains the "sources" for ψ. Let $\mathcal{S}$ be a topological 2-sphere within $\mathcal{R}$ surrounding this source tube. We wish to obtain the charges for ψ by performing a suitable integral over $\mathcal{S}$. One procedure for doing this is to adopt the method (Penrose 1968, 1982, cf. Penrose and Rindler 1986, p75) of *spin-lowering*. For this, we select a (dual) twistor W_α, which for the present purposes means a pair of spinor fields

$$W_\alpha = (\lambda_A, \mu^{A'}) \tag{10}$$

subject to

$$\nabla_{AA'}\mu^{B'} = i\varepsilon_{A'}{}^{B'}\lambda_A,\ \nabla_{AA'}\lambda_B = 0, \tag{11}$$

so $\mu^{B'}$, which depends linearly on position, is a solution of the (conjugate) twistor equation

$$\nabla_A^{(A'}\mu^{B')} = 0. \tag{12}$$

Defining

$$\varphi_{A'B'} = \psi_{A'B'C'}\mu^{C'} \tag{13}$$

we find that, by virtue of (1) and (12), that φ satisfies the self-dual free Maxwell equations

$$\nabla^{AA'}\varphi_{A'B'} = 0,\ \varphi_{A'B'} = \varphi_{(A'B')}, \tag{14}$$

and the *charge* (*electric* + i × *magnetic*) that is assigns to the world-tube is given by

$$\eta = \frac{i}{4\pi}\oint \mathbf{F} \tag{15}$$

where the self-dual 2-form **F** is defined by

$$\mathbf{F} = \varphi_{A'B'}dx_A^{A'} \wedge dx^{AB'} \tag{16}$$

Note that **F** depends linearly on W_α and η depends linearly on **F**. Thus, η's dependence on W_α is given by

$$\eta = Z^\alpha W_\alpha = \omega^A\lambda_A + \pi_{A'}\mu^{A'} \tag{17}$$

for some twistor

$$Z^\alpha = (\omega^A, \pi_{A'}) \tag{18}$$

describing the strength of the charge, for the field ψ, that must be assigned to the world-tube.

We see that the concept of a *twistor* Z^α indeed arises naturally here as the *charge* for a helicity 3/2 massless field. Correpsondingly, a dual twistor would arise as the charge for a helicity $-3/2$ massless field. (Interchange primed and unprimed spinor indices throughout.)

Let us see how the potentials for ψ fit in with this. Note that if

$$\theta^C_{A'} = \gamma^C_{A'B'}\mu^{B'} - \mathrm{i}\rho^{BC}_{A'}\lambda_B \tag{19}$$

then

$$\varphi_{A'B'} = \nabla_{CB'}\theta^C_{A'},\ \nabla^{A'}_B\theta^C_{A'} = 0, \tag{20}$$

by (3), (7), (11) and by (2), (6), (11), and defining the 1-form **A** by

$$\mathbf{A} = \theta_{BB'}\mathrm{d}x^{BB'} \tag{21}$$

we have, from (16),

$$\mathbf{F} = 2\mathrm{d}\mathbf{A} \tag{22}$$

Note that if **A** were defined globally over $\mathscr{S}$, then the definition (15) would yield $\eta = 0$. By (19), (21), this situation always occurs if the second potential ρ, for ψ, is global over $\mathscr{S}$ (since if ρ is global then γ is certainly global), but it will also occur if merely the first potential γ is global provided that W_α belongs to the two-dimensional subspace of dual twistor space given by $\lambda_A = 0$ (see (19)). Referring to (17) we see that the latter situation — where γ is global but ρ is not necessarily so — provides us with the *factor space* of twistor space that is parameterized by the spinor $\pi_{A'}$. I shall refer to this two-dimensional complex vector space as *π-space* and the corresponding *subspace* for which $\pi_{A'} = 0$, parameterized by ω^A, as *ω-space*. We have the short exact sequence

$$0 \longrightarrow \omega\textit{-space} \longrightarrow \textit{twistorspace} \longrightarrow \pi\textit{-space} \longrightarrow 0 \tag{23}$$

(see Penrose and Rindler 1986, p 91) and, by the above (where "global" means global over $\mathscr{S}$):

$$\pi\textit{-space} = \frac{\textit{space of global } \psi s}{\textit{space of global } \gamma s}, \tag{24}$$

$$\omega\textit{-space} = \frac{\textit{space of global } \gamma s}{\textit{space of global } \rho s}. \tag{25}$$

2 Helicity 3/2 fields in Ricci-flat space-times

It was noticed many years ago by Hans Buchdahl (1958), and effectively rediscovered many times subsequently by people interested in supergravity theory etc. (eg. Deser and Zumino 1976), that the (immediate) consistency conditions for a spin 3/2 potential γ given by (2) that arise from the presence of space-time curvature, are automatically satisfied if the trace-free part of the Ricci tensor R_{ab} vanishes. Moreover, the gauge freedom is indeed given by (4), the derivative of a Weyl neutrino field ν as in (5), provided that

$$R_{ab} = 0. \tag{26}$$

It is not clear to me in what sense the description "γ-potential modulo ν-gauge" is fully *consistent* in Ricci-flat space-times, though such claims are frequently made. (In fact the equations normally used by supersymmetry theorists differ in substance and not just in appearance from those given in (2) and (5), in that the symmetry in $A'B'$ is not imposed on γ, and ν is taken to be free with $\nabla_{C(C'}\gamma^C_{A')B'}\varepsilon^{A'B'} = 0$ in addition to $\nabla^{AA'}\gamma^C_{A'B'} = 0$. However, this difference is of no consequence, the local equivalence of the different formalisms being an implication of the local solubility of the inhomogeneous Weyl neutrino equation.) The immediate consistency conditions are obtained by acting on (2) with a contracted derivative $\nabla^{B'}_A$ to obtain

$$\nabla^{B'}_A \nabla^{AA'}\gamma^C_{A'B'} = 0 \tag{27}$$

and on (4) with $\nabla^{BB'}$ to obtain

$$\nabla^{BB'} \nabla^C_{B'}\nu_{A'} = 0, \tag{28}$$

which indeed automatically vanishes when $R_{ab} = 0$, by virtue of the spinor covariant derivative commutators (Penrose and Rindler 1984) and (5). It appears to be the case that all higher-derivative consistency conditions are also automatically satisfied by virtue of (26), but we shall be seeing shortly that the "consistency" that is thereby achieved for the "γ-potential modulo ν-gauge" system is not quite so complete as might have been hoped.

Grounds for some suspicion are already contained in (24), since it would seem that one ought to be able to generalize this relation in order to define the "π-space" for any topological 2-sphere $\mathscr{S}$ in any Ricci-flat space-time $\mathcal{M}$. Although we do not have the ψ-spinors directly, because the defining equation (3) is now not invariant under the gauge transformations (4), we can still provide a concept of "global ψs". For this, we envisage a family of different γs, say $\overset{1}{\gamma}, \overset{2}{\gamma}, \dots, \overset{n}{\gamma}$, with

$$\overset{i}{\gamma} \ defined\ on\ \mathcal{U}_i \tag{29}$$

where $\mathcal{U}_1$, $\mathcal{U}_2, \dots$, $\mathcal{U}_n$ are open sets covering an open region $\mathcal{R}$ of $\mathcal{M}$ containing $\mathcal{S}$. On each (non-vacuous) overlap $\mathcal{U}_i \cap \mathcal{U}_j$ the corresponding γs must be gauge equivalent, i.e. there must be a family of neutrino fields $\overset{ij}{\nu}$ with

$$\overset{i}{\gamma} - \overset{j}{\gamma} = \nabla \overset{ij}{\nu} \ on\ \mathcal{U}_i \cap \mathcal{U}_j \tag{30}$$

(spinor indices suppressed). In $\mathbb{M}$, this is indeed equivalent to having a global ψ in $\mathcal{R}$, but in $\mathcal{M}$, since ψ does not generally exist, one must resort to a piecemeal definition of this kind.

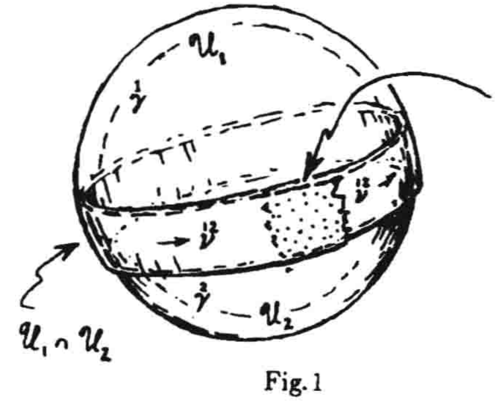
Suspicion naturally arises, however, because if this were to provide us with a definition of π-space for $\mathcal{S}$, we should also have this definition holding unchanged (free of any holonomy) as $\mathcal{S}$ is varied continuously throughout $\mathcal{R}$, and this would seem to provide a "too global" definition of π-space for $\mathcal{M}$. For numerous reasons, a π-space arising in this way is implausible. Indeed, we shall, in a moment, see more directly that fundamental difficulties confront such a global ψ.

Let us return to the situation in $\mathbb{M}$. We can consider covering $\mathcal{R}$ with just two sets $\mathcal{U}_1$, $\mathcal{U}_2$, where we suppose that is just a thickened out S^2 and that $\mathcal{U}_1$ and $\mathcal{U}_2$ are thickened out northern and southern hemispheres. There is an annular intersection region $\mathcal{U}_1 \cap \mathcal{U}_2$ (a thickened out equator) along which $\overset{1}{\gamma} - \overset{2}{\gamma}$ is pure gauge:

$$\overset{1}{\gamma} - \overset{2}{\gamma} = \nabla \overset{12}{\nu}. \tag{31}$$

However, if the "π-charge" is to be non-zero, we find that $\overset{12}{\nu}$ is not single-valued but it must jump by a constant spinor field as sone proceeds once around the annulus, this spinor field being a numerical multiple of (in fact $(2\pi \mathrm{i})^{-1}\times$) the actual value $\pi_{A'}$ of the "π-charge" of the spin 3/2 field in question (see Fig 1).

It is important to note that the reason this works is that *constant* spinors $\nu_{A'}$ do not show up in the gauge transformations (4), since then $\nabla^{C}_{A'}\nu_{B'} = 0$. These constants correspond to the "gauge transformations of the second kind" that are



$\overset{12}{\nu}$ jumps by
a constant
spinor that is
proportional to
$\pi_{A'}$

Fig. 1

familiar in electromagnetic theory. In fact the procedure just outlined is effectively an example of "evaluation" of a cohomology class, although the covering $\mathcal{U}_i$ should have been chosen more finely than this for the procedure strictly to apply: all the open sets and their various multiple intersections should be topologically trivial (unlike the annular region $\mathcal{U}_1 \cap \mathcal{U}_2$). If it is assumed that these multiple intersections are indeed topologically trivial, the evaluation procedure is as follows: we have

$$\mathbf{F} = 2\mathrm{d}\,\overset{i}{\mathbf{A}},\ \overset{i}{\mathbf{A}}\ \text{defined on } \mathcal{U}_i. \tag{32}$$

Also

$$\overset{i}{\mathbf{A}} - \overset{j}{\mathbf{A}} = \mathrm{d}\overset{ij}{\nu},\ \overset{ij}{\nu}\ \text{defined on } \mathcal{U}_i \cap \mathcal{U}_j; \tag{33}$$

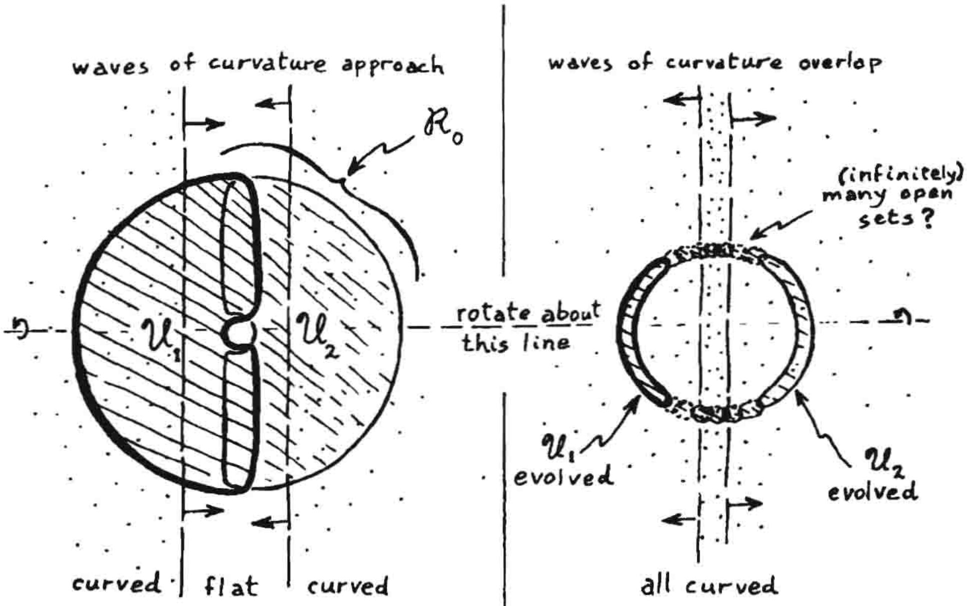
$$\overset{ij}{\nu} + \overset{jk}{\nu} + \overset{ki}{\nu} = \overset{ijk}{\Pi} = const.\ defined\ on\ \mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k \tag{34}$$

(The fact that the $\overset{ijk}{\Pi}$ are constant follows from the above, since $\mathrm{d}\overset{ijk}{\Pi} = 0$.) The $\overset{ijk}{\Pi}$ $(=\overset{ijk}{\Pi}_{A'})$ provide a 2-cocycle that is evaluated against the 2-cycle that is the 2-sphere $\mathcal{S}$ itself. (This amounts to adding together the different $\overset{ijk}{\Pi}$ with appropriate signs and possible redundancies.) The result is $((2\pi i)^{-1}\times)$ the required "π-charge" $\pi_{A'}$.

Suppose, now, that a slight perturbation of the space-time $\mathbb{M}$ is made so that it becomes a general curved space-time $\mathcal{M}$. The whole procedure must now break down because constant νs do not generaly exist. The π-charge, in these circumstances must necessarily turn out to be zero!

There is something very puzzling about this since it would have seemed that one "ought" to be able to piece together a family of γ-potentials for a covering of $\mathcal{R}$ in $\mathcal{M}$ that is close to the original family (with a non-zero π-charge) in $\mathbb{M}$. The above argument shows that this intuition is incorrect, however. In fact the situation is even more puzzling that just this. We can envisage a space-time $\mathcal{M}$ containing an initial flat region within which one can define a family of γ-potentials with non-zero total π-charge. As the (Ricci-flat) space-time evolves, the domain of dependence of the region $\mathcal{R}$, in which the γs are initially defined, becomes entirely within a curved region of $\mathcal{M}$, and a π-charged family of γ-potentials becomes impossible. Such a situation is illustrated in Fig 2. Here we have a thick "plate" of flat space-time that is being encroached upon by two (plane) wave-fronts, behind which are two regions of (generic) curvature. Before the waves collide, there is a region $\mathcal{R}_0 = \mathcal{U}_1 \cup \mathcal{U}_2$ (a **thick shell, which is a very much thickened out 2-sphere) that has a sufficiently large part of it in the flat region, so that a non-zero π-charged pair of γs can be defined. It would *seem* that there is no obstruction to extending these γs into the whole of $\mathcal{U}_1$ and $\mathcal{U}_2$, as in Fig. 2. As the waves come together and then separate again, having passed through one another, the space-time is entirely curved; yet the domain of dependence of $\mathcal{R}_0$, though as time passes it thins down considerably from the "thick" region $\mathcal{R}_0$, is still extended enough that the π-charge can be defined — and it should be non-zero — *provided* that a sufficient family of evolved γs exists.**

It should be reasonably clear that *if* the γs can be evolved adequately in this way, then we shall be presented with a contradiction. The π-charge cannot simply disappear the moment that an (arbitrarily small) amount of curvature is encountered. But



Waves of curvature coming together: non-zero π-charge still possible

Waves of curvature separating: π-charge must vanish; domain of dependence now a thin sphere-like shell

Fig. 2

what goes wrong with the γ-evolution? The answer seems to be that within the domain of dependence of each $\mathcal{U}_i(\cap \mathcal{R}_0)$ the relevant $\overset{i}{\gamma}$ does indeed evolve consistently. However, as time evolves, these domains of dependence pull away from each other and the needed intersection regions need no longer exist. One might anticipate that new open sets could be continually introduced to cover these intersections before they disappear, but these new sets, in turn, would pull away from each other. In order that this procedure can be continued sufficiently to cover the whole of the domain of dependence of $\mathcal{R}_0$, it would be necessary that the new open sets that are introduced can be chosen to be sufficiently "large". The obstruction to the size of such an open set, on which the corresponding γ can be globally defined, would be something of the nature of a "radius of convergence". However, in the situation under consideration, the space-time $\mathcal{M}$ is necessarily *non-analytic* (though it can be C^∞ and piecewise analytic), and the γs would be correpsondingly non-analytic. As the waves come together, the relevant radii of curvature may be reduced to zero and the covering ceases to exist, yielding a fundamental obstruction to the continuation of the "γ- potential modulo ν-gauge" spin 3/2 field within its own domain of dependence, despite the fact that in an *analytic* Ricci-flat space-time this description appears to be "consistent".

To some extent, the above description is merely an indirect "inferrence" concerning what presumably "must" happen in this puzzling and almost paradoxical situation. It would be important to gain some more direct insights into what is actually going on. (There may be some connection with the "patching problems" arising for the equations of motion of monopoles in non-Abelian gauge fields; cf. Chan, Scharbach and Tsou 1986.)

It is clear that some new insights are required if these ideas are to lead to a satisfactory "twistor" concept for general Ricci-flat space-times. it is necessary, for example, that any such concept provide a *non-linear* space of twistors, whereas the ideas, as they stand, have too much of a linear flavour about them. We need non-linearity because, in particular, we would wish to reproduce the curved twistor space of the "non-linear graviton construction" (Penrose 1976) in the case that the space-time $\mathcal{M}$ (now complex) is anti-self-dual ($\tilde{\Psi}_{A'B'C'D'} = 0$), in the notation of Penrose and Rindler 1986, p 129) and, as with that construction, we would eventually hope to be able to reconstruct the space-time geometry of a general Ricci-flat space-time from the (non-vector space) complex structure of its twistor space.

In a general way, I feel that a certain amount of optimism is in order, and to end on an encouraging note, I should point out that in a complex vacuum space time that is anti-self-dual ($\tilde{\Psi}_{A'B'C'D'} = 0$) the spin 3/2 *field* quantity $\psi_{A'B'C'}$ can be defined according to (3) and in this situation it is gauge-invariant and satisfies (1). There is now no obstruction to defining global ψ-fields, with non-zero π-charge and this is consistent with the fact that constant spinors $\Pi_{A'}$ *do* exist, there being a global π-space that can indeed be defined by (24). There is some hope that the twistor space defined by the α-planes in $\mathcal{M}$ (Penrose 1976) may be able to be reconstructed.

At the other end of the scale, we can consider *self*-dual Ricci-flat complex space-times ($\Psi_{ABCD} = 0$), and here it is the second potential ρ, together with its gauge quantity χ that can be consistently defined. Now it is the definition (25) that makes sense and it is the ω-space (with $\pi = 0$) that is globally defined for the space-time. To understand the *general* case (neither self-dual not anti-self-dual), however, much more work is needed.

Acknowledgments

I thank Rob Baston, Florence Tsou and particularly Andrew Hodges for stimulating discussions.

References

1. Buchdahl, H.A. (1958) On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field *Nuovo Cim.* 10, 96-103

2. Chan, H.-M., Scharbach, P. and Tsou, S.T. (1986) Action principle and equations of motion for nonabelian monopoles *Annals of Physics*, 167, 454-72

3. Deser, S. and Zumino, B. (1976) Consistent supergravity *Phys. Lett.* 62B, 335-7

4. Fierz, M. (1940) Uber den Drehimpuls von Teilchen mit Ruhemasse null und beliebigen Spin *Helv. Phys. Acta.* 13, 45-60

5. Hodges, A.P. (1990a) Elemental states in *Further Advances in Twistor Theory Vol 1*: (L.J. Mason and L.P. Hughston, eds. Pitman Notes in Mathematics 231, Longman, Harlow, Essex, UK and John Wiley, New York, USA), pp 308-310 (From Twistor Newsletter 22 (1986))

6. Hodges, A.P. (1990b) Twistor diagrams and Feynman diagrams, in *Twistors in Mathematics and Physics* (T.N. Bailey and R.J. Baston, eds. Cambridge Univ. Press, Cambridge), pp 339-366

7. Penrose, R. (1968) Twistor quantization and curved space-time *Int. J. Theor. Phys.* 1, 61-99
8. Penrose, R. (1976) Nonlinear gravitons and curved twistor theory, *Gen. Rel. Grav.* 7, 31-52
9. Penrose, R. (1980) A brief outline of twistor theory, in *Cosmology and Gravitation Supergravity*, eds. P.G. Bergmann and V. de Sabbata (Plenum, New York)
10. Penrose, R. (1982) Quasi-local mass in general relativity, *Proc. Roy. Soc. London* A381, 53-63
11. Penrose, R. (1990) Twistor theory for vacuum space-times: a new approach, *Twistor Newsletter* 31, 6-8
12. Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time*, Vol 1: *Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge)
13. Penrose, R. and Rindler, W. (1986) *Spinors and Space-Time*, Vol 2: *Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge)
14. Sachs, R.K. and Bergmann, P.G. (1958) Structure of particles in linearized gravitational theory, *Phys. Rev.* 112, 674-80

203

TWISTORS AS CHARGES FOR SPIN 3/2 IN VACUUM

Twistor Newsletter **32** (1991). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

See Chapter 202. It seems, in view of the special example presented here, that there is a subtlety that remains to be understood.

134 Twistors as charges for spin 3/2 in vacuum. *Twistor Newsletter* **32** (1991). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001

§III.4.20 **Twistors as Charges for Spin $\frac{3}{2}$ in Vacuum** *by R. Penrose* ($\mathbb{TN}$ 32, March 1991)

In §III.4.19, I put forward the suggestion that it might be possible to develop a twistor theory for general vacuum space-times by describing a twistor as a kind of helicity

Twistors as charges for spin 3/2 in vacuum. *Twistor Newsletter* **32** (1991). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001

$\frac{3}{2}$ *elemental state.* Elemental states were introduced by Andrew Hodges in §I.5.11. They provide a cohomological space-time interpretation of the internal lines $\circ\!\!-\!\!\!-\!\!\bullet$ that make up twistor diagrams. For spin $\frac{3}{2}$, the elemental states would correspond to the internal lines $\circ\!\overset{-2}{-\!\!\!-}\!\bullet$ that occur with perturbative descriptions of gravitational scatterings, regarding $w\overset{-2}{-\!\!\!-}z$ as a 'twistor function' $f(w_\alpha)$ of homogeneity degree -1 describing a twistor Z^α as some kind of cohomological space-time field (of helicity $\frac{3}{2}$, since $f(W_\alpha)$ has homogeneity $1 = 2s - 2 = 2 \times \frac{3}{2} - 2$). The idea here was to use the fact that spin $\frac{3}{2}$ fields are, in an appropriate sense 'consistent' in (and only in) Ricci-flat space-time (see Buchdahl 1958, Deser and Zumino 1976, and especially Julia 1982), so the interpretation of twistors as helicity $\frac{3}{2}$ elemental states might also carry over to general Ricci-flat space-times. An essential problem had seemed to be how to characterize the particular degree 1 'twistor function $f(W_\alpha)$', in terms of its cohomological helicity $\frac{3}{2}$ space-time field that actually picks out $W\overset{-2}{-\!\!\!-}Z$ amongst all twistor functions $f_{-2}(W_\alpha)$. However, in flat space there is a special cohomological role for $W\overset{-2}{-\!\!\!-}Z$, which can be reinterpreted in space-time terms, in which Z^α plays a role as a ('disembodied') *charge* for a helicity $\frac{3}{2}$ field.

There are various different formulations of the massless spin $\frac{3}{2}$ equations. The one I used in §III.4.19 might be called the *'Dirac form'* where, in flat space-time $\mathbb{M}$ we have the chain

$$\begin{aligned}
&\Psi_{A'B'C'} = \Psi_{(A'B'C')}, \qquad \nabla^{AA'}\Psi_{A'B'C'} = 0\\
&\gamma^{A}_{B'C'} = \gamma^{A}_{(B'C')}, \qquad \nabla^{BB'}\gamma^{A}_{B'C'} = 0, \qquad \nabla_{AA'}\gamma^{A}_{B'C'} = \Psi_{A'B'C'}\\
&\rho^{AB}_{C'} = \rho^{(AB)}_{C'}, \qquad \nabla^{CC'}\rho^{AB}_{C'} = 0, \qquad \nabla_{BB'}\rho^{AB}_{C'} = \gamma^{A}_{B'C'}
\end{aligned}$$

with gauge freedom

$$\begin{aligned}
&\gamma^{A}_{B'C'} \mapsto \gamma^{A}_{B'C'} + \nabla^{A}_{B'}\nu_{C'} \quad \text{with} \quad \nabla^{AC'}\nu_{C'} = 0\\
&\rho^{AB}_{C'} \mapsto \rho^{AB}_{C'} + \varepsilon^{AB}\nu_{C'} + \mathrm{i}\,\nabla^{A}_{C'}\chi^{B} \quad \text{with} \quad \nabla_{AC'}\chi^{A} = 2\,\mathrm{i}\,\nu_{C'}\,.
\end{aligned}$$

There is also a 'gauge freedom of the second kind' (as in EM-theory) which leaves the potentials unchanged, given by

$$\nu_{A'} = \Pi_{A'}, \quad \chi^A = \Omega^A$$

where $(\Omega^A, \Pi_{A'})$ are the spinor parts of some twistor, i.e.,

$$\nabla_{AA'}\Omega^B = -\,\mathrm{i}\,\epsilon_A{}^B\Pi_{A'} \quad \text{with } \Pi_{A'} \text{ constant.}$$

136 Twistors as charges for spin 3/2 in vacuum. *Twistor Newsletter* **32** (1991). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001

We have an exact sequence

$$\begin{array}{ccccccccc} 0 & \rightarrow & \text{gauge potentials of } 2^{nd} \text{ kind} & \rightarrow & \text{gauge quantities} & \rightarrow & \text{potential} & \rightarrow & \text{fields} & \rightarrow & 0\,. \\ & & (\Omega,\Pi) & & (\chi,\nu) & & (\rho,\gamma) & & (\psi) & & \end{array}$$

In a general Ricci-flat curved space-time $\mathcal{M}$, most of this does not carry through, but we can still retain (Dirac form)

$$\gamma^A_{B'C'} = \gamma^A_{(B'C')},\ \nabla^{BB'}\gamma^A_{B'C'} = 0,\ \text{with gauge } \nabla^A_{B'}\nu_{C'},\ \text{where } \nabla^{AC'}\nu_{C'} = 0$$

as a consistent system (subject to some puzzling phenomena that we shall need to come back to later). (It is easy to see that the Ricci-flat condition is necessary and sufficient for the gauge freedom to be actually $\nabla^A_{B'}\nu_{C'}$; that it is sufficient also for the γ equations is more complicated.) People who work with supersymmetry, however, prefer what one may call the *'Rarita-Schwinger form'* where we do *not* impose symmetry $\gamma^A_{B'C'} = \gamma^A_{(B'C')}$, and take our equations as $\varepsilon^{B'C'}\nabla_{A(A'}\gamma^A_{B')C'} = 0$, $\nabla^{B'(B}\gamma^{A)}_{B'C'} = 0$,

with gauge $\nabla^A_{B'}\nu_{C'}$, where $\nu_{C'}$ is unrestricted.

The first of these equations asserts the total symmetry of $\Psi_{A'B'C'} = \nabla_{A(A'}\gamma^A_{B')C'}$; however, it should be noted that if $\mathcal{M}$ is not A.S.D. (anti-self-dual), then $\Psi_{\ldots}$ is not gauge-invariant (in either the Dirac or R.-S. form). The R.- S. form is easily reduced to the Dirac form by using a gauge choice for ν (solving an inhomogeneous Weyl neutrino equation) which makes γ symmetric, the remaining gauge freedom being subject to $\nabla^{AA'}\nu_{A'} = 0$.

In $\mathbb{M}$, we may obtain the charges for a helicity $\frac{3}{2}$ field by spin-lowering. Take the spinor field $\mu^{A'}$ to be the primary part of a dual twistor $w_\alpha = (\lambda_A, \mu^{A'})$:

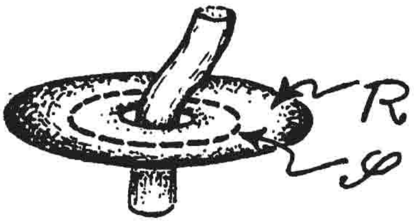
$$\nabla_{AA'}\mu^{B'} = \mathrm{i}\,\varepsilon'_{A}{}^{B'}\lambda_A\,, \qquad \lambda_{A'}\,\text{const}$$

and set

$$\varphi_{A'B'} = \Psi_{A'B'C'}\mu^{C'}$$

for some Ψ in an open region $\mathcal{R}$ surrounding a world tube.

Twistors as charges for spin 3/2 in vacuum. *Twistor Newsletter* **32** (1991). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001



We can use the self-dual Maxwell field φ in a charge integral over some 2-surface $\mathscr{S}$ in $\mathcal{R}$ surrounding the tube:

$$\frac{\mathrm{i}}{4\pi}\oint \varphi_{A'B'}dx_{A}^{A'} \wedge dx^{AB'} = \zeta(w) = \text{ electric } + \mathrm{i} \times \text{ magnetic}$$

where $\zeta(w)$ is linear in w_α and therefore $\zeta(w) = Z^\alpha W_\alpha$ for some Z^α. This Z^α is the required *charge* for Ψ. Note that it does not depend on the particular choice of $\mathscr{S}$.

We want to do something here that has a chance of holding in $\mathcal{M}$ and hence of providing a definition of a twistor in general Ricci-flat space-times (and independent of any particular choice of $\mathscr{S}$). A possible suggestion would be to generalize the following, which works in $\mathbb{M}$. Consider fields $\Psi_{\dots}$ which are global in $\mathcal{R}$ ($\subset \mathbb{M}$). We shall say that two such fields are *equivalent* if their difference has a (second) potential ρ (and therefore also a potential γ) which is global in $\mathcal{R}$. The space $\mathbb{T}^\alpha$ of twistors can be identified with this space of equivalence classes

$$\mathbb{T}^\alpha = \{\text{global } \Psi_s\}/\{\text{global } \rho_s\}.$$

We can also find the space $\mathbb{S}_{A'}$ of constant $\pi_{A'}$-spinors, and the space $\mathbb{S}^A$ of constant ω^A-spinors when $\pi_{A'} = 0$ as, respectively:

$$\mathbb{S}_{A'} = \{\text{global } \Psi_s\}/\{\text{global } \gamma_s\}, \quad \mathbb{S}^A = \{\text{global } \gamma_s\}/\{\text{global } \rho_s\}$$

illustrating

$$0 \to \mathbb{S}^A \to \mathbb{T}^\alpha \to \mathbb{S}_{A'} \to 0.$$

How much of this generalizes for Ricci-flat $\mathcal{M}$? It may be remarked that if $\mathcal{M}$ is A.S.D. then Ψ-fields exist and have the same relation to γs as in $\mathbb{M}$. This is consistent with the above and with the fact that $\mathbb{S}_{A'}$ is well-defined in the A.S.D. case. If $\mathcal{M}$ is S.D.

then ρ-potentials exist (locally) and have the same relation to γs as in $\mathbb{M}$, consistently with the fact that $\mathbb{S}^A$ is well-defined in the S.D. case. In the A.S.D. case we can hope to go further and obtain the full curved twistor space $\mathscr{T}$ of the 'nonlinear graviton' construction. Although we do not have ρs locally, we can weaken the equation relating ρ to γ to

$$\gamma^{A}_{B'C'} = \nabla_{B(B'}\hat{\rho}^{AB}_{C')},$$

since this should be locally soluble, and then state the equivalence between two Ψs in terms of the existence of a complex 2-surface (the α-surface of 'Z') on which their respective $\hat{\rho}$s have an appropriate globality. This criterion seems too clumsy to be what is ultimately needed, but something of a nonlinear nature is certainly required.

If $\mathcal{M}$ is general (Ricci-flat), or even merely S.D., we need a concept of a 'global Ψ' throughout $\mathcal{R}$ $(\subset \mathcal{M})$, even though the actual spinor Ψ is not gauge invariants. In the case of $\mathbb{M}$, we can state such globality entirely in terms of γ by saying that $\mathcal{R}$ can be covered by open sets $\{\mathcal{U}_i\}$ where $\exists\ \overset{i}{\gamma}$ in $\mathcal{U}_i$ with

$$\overset{i}{\gamma} - \overset{j}{\gamma} = \nabla\ \overset{ij}{\nu} \ \text{ in each (non-empty) } \mathcal{U}_i \cap \mathcal{U}_j$$

where

$$\overset{ij}{\nu} \ \text{ exists in } \mathcal{U}_i \cap \mathcal{U}_j$$

with $\nabla^{AC'}\ \overset{ij}{\nu}_{C'} = 0$ (taking the Dirac form). We can then define

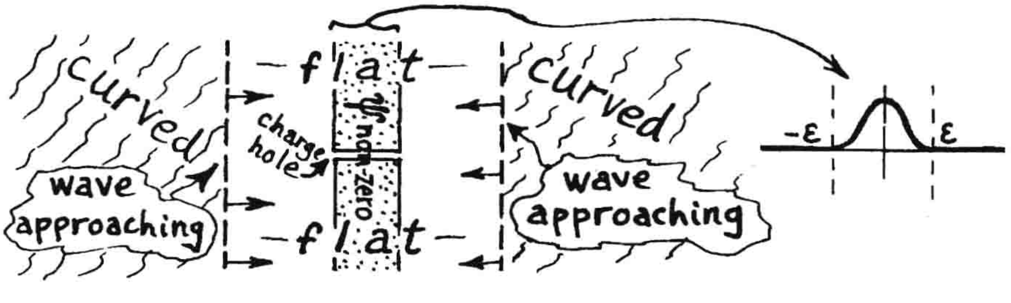
$$\overset{ijk}{\Pi} = \overset{ij}{\nu} + \overset{jk}{\nu} + \overset{ki}{\nu} \ \text{ in } \mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k$$

and note that the Πs are constant since $\nabla\ \overset{ijk}{\Pi} = 0$. (We could also do the same with the ρs, χs and Ωs to obtain a twistor $(\overset{ijk}{\Omega^A}, \overset{ijk}{\Pi}_{A'})$ in $\mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k$.) Appropriately adding these up, we obtain $(\frac{1}{2\pi\mathrm{i}}\times)\pi_{A'}$ (or $\frac{1}{2\pi\mathrm{i}}(\omega^A, \pi_{A'})$) as required. This is an example of an 'evaluation procedure' for a cohomology class — here an H^2 element, as we move 3 steps back from ψ in our exact sequence of potentials in order to get the charge for ψ.

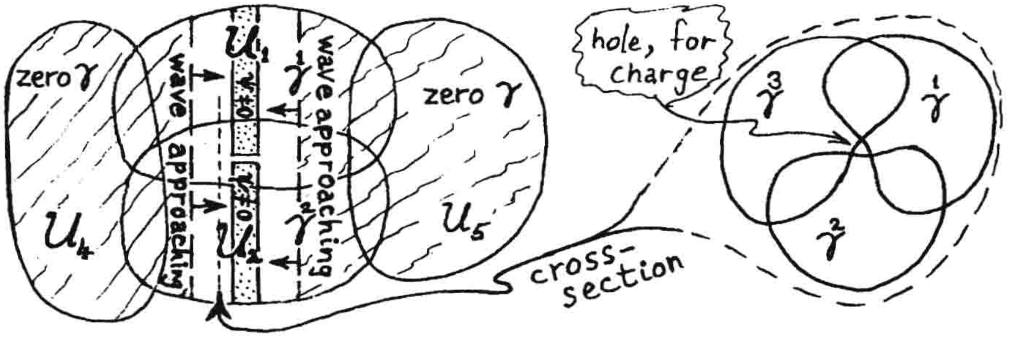
In a (non-A.S.D.) $\mathcal{M}$, something must go wrong with our original γs if we are to expect a non-zero $\pi_{A'}$-charge. For there are *no* constant $\Pi_{A'}$ spinors other than zero. This seems to be an almost paradoxical situation, because we can envisage situations where a 'ψ' with a non-zero π-charge is set up in an initially flat region of space-time, and this 'ψ' evolves, within its domain of dependence, into a full region of non-zero curvature. (Here 'ψ' means γs modulo $\nabla\nu$s in a covering of $\mathcal{R}$, as described above.) We can arrange that the initial ψ-field is non-zero entirely within a thin slab, and singular only just around the centre of the slab, e.g., of the form

$$\zeta^{-1}o_{A'}o_{B'}o_{C'}B(u) \quad \text{or} \quad \zeta^{-1}\iota_{A'}\iota_{B'}\iota_{C'}B(v)$$

at $u+v=0$ ('$t=0$') in a flat region with standard null coordinates $(u,v,\zeta,\bar{\zeta})$ and constant dyad (o^A,ι^A) for $\mathbb{M}$ (metric $ds^2 = 2dudv - 2d\zeta d\bar{\zeta}$, $\frac{\partial}{\partial v} = o^A\bar{o}^{A'} = l^a$, $m^a = \frac{\partial}{\partial\zeta}$, $\frac{\partial}{\partial u} = \iota^A\bar{\iota}^{A'} = n^a$) where B is a C^∞ 'bump function' non-zero only in $(-\varepsilon,\varepsilon)$. Plane waves approach this region from either side:

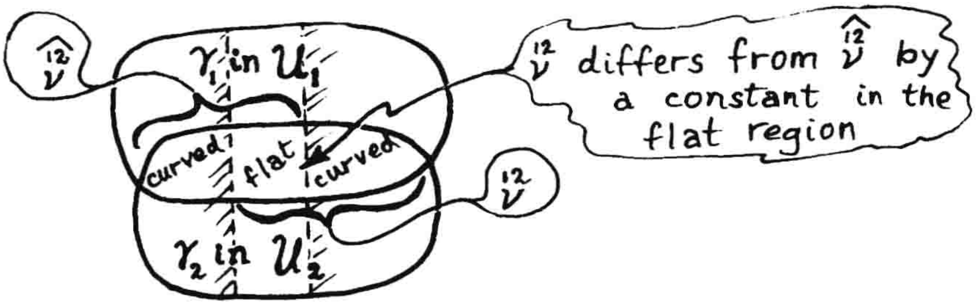

curved
flat
charge hole
ψ non-zero
flat
curved
wave approaching
wave approaching
-ε
ε

It would *seem* that we could cover a huge region of $\mathcal{M}$ with large open sets with large overlaps — large enough so that the domains of dependence of these open sets, individually, still overlap after the waves have collided and the space-time is entirely curved.

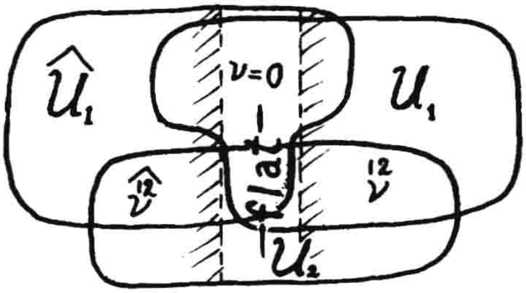

zero γ
wave approaching
𝒰₁
ψ≠0
γ¹
wave approaching
zero γ
ψ≠0
γ²
𝒰₄
𝒰₂
𝒰₅
hole, for charge
γ³
γ¹
γ²
cross-section

Plane waves have constant spinors but we can tilt them very slightly so that these constants are not in the directions of o^A or ι^A. Thus the $\pi_{A'}$-charges must disappear (these being in the directions of $o_{A'}$ or $\iota_{A'}$). How can this happen?

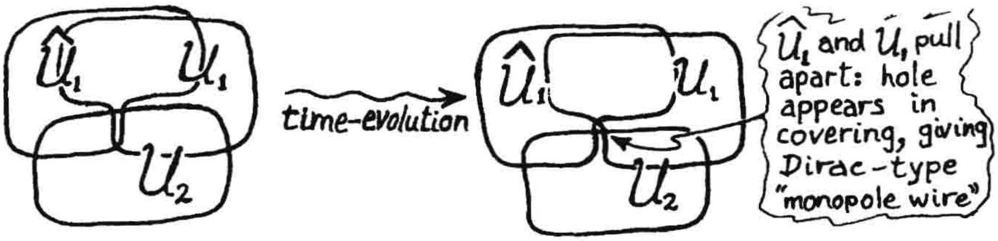
What actually occurs in this situation is somewhat surprising. It turns out that we find pairs of γs that are 'locally' but not properly globally gauge equivalent:


$\hat{\overset{12}{\nu}}$
γ_1 in $\mathcal{U}_1$
curved
flat
curved
γ_2 in $\mathcal{U}_2$
$\overset{12}{\nu}$
$\overset{12}{\nu}$ differs from $\hat{\overset{12}{\nu}}$ by a constant in the flat region

To make this legitimate we must subdivide, say to


$\hat{\mathcal{U}}_1$
$\nu=0$
$\mathcal{U}_1$
flat
$\hat{\overset{12}{\nu}}$
$\overset{12}{\nu}$
$\mathcal{U}_2$

where $\hat{\overset{1}{\gamma}}$ and $\overset{1}{\gamma}$ are actually *identical* in $\hat{\mathcal{U}}_1 \cap \mathcal{U}_1$! Here, all three of $\hat{\overset{1}{\gamma}}$, $\overset{1}{\gamma}$ and $\overset{2}{\gamma}$ *are* properly gauge equivalent on their overlaps. However, the needed region of triple overlap now *disappears* as the waves collide:


$\hat{\mathcal{U}}_1$
$\mathcal{U}_1$
$\mathcal{U}_2$
time-evolution
$\hat{\mathcal{U}}_1$
$\mathcal{U}_1$
$\mathcal{U}_2$
$\hat{\mathcal{U}}_1$ and $\mathcal{U}_1$ pull apart: hole appears in covering, giving Dirac-type "monopole wire"

It seems that the twistor-charge 'escapes' along a Dirac-type 'monopole wire'. Apparently we must resort to such 'wire-type' descriptions (with non-global γs) if we are to recapture our twistor concept. Perhaps this is not a bad thing in the long run, since a *nonlinear* (non-cohomological, in the strict sense) concept is needed.

Work is in progress. There are further relations to twistor diagrams and I hope that they will be reported on later.

Especial thanks go to Paul Tod, who worked out the actual propagation of the above charged fields and potentials into plane-wave backgrounds—showing that we do not get a Hans Lewy obstruction (at least at this stage) as had seemed possible; also to Lionel Mason for putting me straight on the R.-S. equations. Comments from Andrew Hodges, Rob Baston and Florence Tsou have also been valuable.

References

Buchdahl, H.A. (1958) On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field, Nuovo Cimento, **10**, 335–7.

Deser, S. & Zumino, B. (1976) Consistent supergravity, Phys. Lett., **62B**, no. 4, 335–7.

Julia, B. (1982) Linear System associated with Einstein's equations, Comptes Rendus Acad. Sci. Paris, Sér. **II**, no. 2, 113–6.

204

THE MASS OF THE CLASSICAL VACUUM

The philosophy of vacuum. Eds S. Saunders and H. R. Brown, 21–6. Clarendon Press, Oxford, 1991

This article points out difficulties confronting the very notion of 'matter' according to modern classical physics.

The mass of the classical vacuum. *The Philosophy of Vacuum.* Eds S. Saunders and H. R. Brown, 21–6. 1991

2

The Mass of the Classical Vacuum

R. PENROSE

There is something a bit paradoxical in the lessons classical physics has to teach us about the physical nature of matter. We may ask, What indeed *is* 'matter'? The commonsense reply might be that it is the real substance of which actual physical objects—the 'things' of this world—are composed. It is what you, I, and our houses are made of. How, then, does one actually quantify this substance? Our elementary physics textbooks provide us with Newton's clear answer: it is the *mass* of an object, or system of objects, that measures the quantity of matter that it contains. This, indeed, now seems right; there is no other physical quantity that can seriously compete with mass as the true measure of total substance. Moreover, it is *conserved*—so that the mass of any system whatever must be unchanging with time.

Yet Einstein's famous (1905) formula from special relativity,

$$E = mc^2,$$

tells us that mass and energy are interchangeable with one another. Mass is still conserved, but now it seems less clearly to be the true measure of actual substance. Energy, after all, depends upon the speed with which that substance is seen to be travelling. The energy of motion in an express train is considerable, but if we happen to be sitting in that train, then, according to our own reference frame, the train possesses no motion at all. The energy of that motion (though not the heat energy of the individual particles, nor the rest-mass energy of those particles) is now reduced to zero by our particular choice of frame. The total mass of the express train, being proportional to its energy, appears to be less to a traveller on the train than to someone who remains stationary on the ground as the train speeds by.

© R. Penrose 1991

This article is based largely on a passage (the final section of ch. 5) from *The Emperor's New Mind*, by R. Penrose (Oxford University Press, 1989).

The mass of the classical vacuum. *The Philosophy of Vacuum.* Eds S. Saunders and H. R. Brown, 21–6. 1991

Here the discrepancy in the total mass remains relatively small because the train's speed is small compared with the speed of light. On the other hand, for a striking example where the effect of Einstein's mass–energy relation is at its most extreme, consider a π^0-meson. It is certainly a material particle, with a well-defined (positive) mass. After about 10^{-16} of a second, it decays—almost certainly into a pair of photons. In the frame of the π^0-meson, each photon carries away half the energy and, indeed, half of the π^0-meson's mass. Yet this mass is of the nebulous kind: pure energy. For if we were to travel rapidly in the direction of one or other of the photons, then we could reduce the mass–energy of that photon to as small a value as we please.

All this makes for a consistent picture of conserved mass, but it is not quite the one that we had before. Mass can still, in a sense, measure 'quantity of matter', but there has been a distinct change of viewpoint. The conserved quantity that takes over the role of mass is now the entire energy-momentum 4-vector. The time component,

$$p^0 = E/c = m_I c,$$

describes only the observer-dependent 'inertial mass' m_I (or, equivalently, the energy) of a particle or system relative to the observer.

This 4-vector has a (non-negative) Minkowskian norm m_R, defined by

$$c^2(m_R)^2 = E^2c^{-2} - \mathbf{p}\cdot\mathbf{p} = (p^0)^2 - (p^1)^2 - (p^2)^2 - (p^3)^2,$$

where $\mathbf{p} = (p^1, p^2, p^3)$ is the 3-momentum. The norm is the rest-mass m_R (or, equivalently, the rest energy $m_R c^2$). One might try to take the view that m_R would be a good measure of 'quantity of matter', since this is observer-independent. However, it is not additive: if a system splits into two, then the original rest-mass is not the sum of the resulting two rest-masses. Recall the π^0-meson decay considered above. The π^0-meson's rest-mass is non-zero, while the rest-masses of each of the two resulting photons is zero. (Any particle that moves with the speed of light must have rest-mass zero.) The additivity property, now in the sense of vector addition, holds for the entire 4-vector, and this vector must thus be our measure of 'quantity of matter'.

Think now of Maxwell's electromagnetic field. As Maxwell himself clearly pointed out (1865, 1873), this field carries energy. Thus, by $E = mc^2$, the field must also have mass. Thus, Maxwell's field is also matter! This must now certainly be accepted, since Maxwell's field is intimately involved in the forces that bind particles together to form atoms. There must be a substantial contribution to any body's mass from the electromagnetic fields within it. (Though, unfortunately, this contribution is incalculable, on present theory—which gives *infinity* as its provisional, but unhelpful, answer!)

The energy in an electromagnetic field can be described as an energy *density*, i.e. as energy per unit volume, which is the normal way of describing the energy for a continuous medium. It is the spatial *integral* of this density that provides the total energy of the system. The energy density is a component not of a 4-vector, but of a valence-2 tensor. If there are many continuous media present (e.g. quantum field descriptions of particles), then we have an energy density, and hence a corresponding tensor, for each one. We add all these tensors together to obtain a quantity

$$T_{ab},$$

referred to as the *energy momentum tensor* of the system.

What about Einstein's gravitational field? In many ways it resembles Maxwell's. As with Maxwell's theory, bodies in motion can emit waves, and like electromagnetic waves they travel with the speed of light and carry energy. Yet this energy is not measured in the standard way, which would be by the above energy momentum tensor. In Einstein's equation

$$R_{ab}-\tfrac{1}{2}Rg_{ab}=-8\pi GT_{ab}$$

(where G is Newton's gravitational constant, R_{ab} is the Ricci tensor, g_{ab} the metric tensor, and R the scalar curvature), the T_{ab} on the right is supposed to be describing the entire non-gravitational energy. In vacuum, which in Einstein's theory means in the absence of all physical fields except gravity, the energy-momentum tensor is zero, whence $R_{ab}=0$; but there can still be a gravitational field present. This gravitational (tidal distortion) field is described by the full *Riemann curvature tensor* R_{abcd}, which has a total of 20 components. The Ricci tensor has just ten components, and the remaining ten collect together in the form of another tensor,

$$C_{abcd},$$

called the *Weyl tensor*. In vacuum, such as inside a (pure) gravitational wave, the Weyl tensor still survives, and it can be thought of as describing the free gravitational field.

The Weyl tensor is not an appropriate object for directly describing gravitational energy, however. (It has too many indices, for example.) Nevertheless, gravity *does* contribute to the total mass–energy of a physical system. The simplest way of seeing this is to consider two masses. When they are far apart, the total energy of the system is somewhat greater than when they are close together, owing to the Newtonian potential energy contribution. Thus, by $E=mc^2$ they must have a slightly greater mass when they are far apart than when they are close together. The difference would have to come from the gravitational field in the space between the masses. But this cannot arise as an integral of the energy density locally defined in T_{ab} because that energy density is zero outside the masses. Also, as mentioned above, (pure) gravitational waves carry energy, yet the energy density throughout the waves is everywhere zero.

These problems are related to the fact that the 'conservation law'

$$\nabla^a T_{ab}=0$$

that is enjoyed by the energy-momentum tensor is a 'covariant' one (∇^a denoting covariant derivative), and does not give rise to the integral conservation law that one would like, namely one asserting that the total energy of a physical system is actually constant. In a well-known attempt to resolve these issues, Einstein introduced a quantity $\mathfrak{T}_{ab}$, referred to as the energy momentum *pseudo*-tensor, which was intended to take the energy of the gravitational field into account. This did give rise to an integral conservation law, but it suffered from the very serious drawback of depending heavily on the particular system of coordinates that happened to have been chosen for the problem at hand. The components of $\mathfrak{T}_{ab}$ therefore had no local physical meaning (a difficulty that was already appreciated by Einstein), and one certainly cannot take this pseudo-tensor description as providing the 'true' measure of the mass–energy distribution in the gravitational field.

One might take the view, nevertheless, that somehow the *curvature* of space–time—as measured by the surviving Weyl components of the curvature tensor—can still represent the 'stuff' of gravitational waves. But, as is indicated by the above arguments, gravitational energy is *non-local*, which is to say that one cannot determine what the measure of this energy is by merely examining the curvature of space–time in limited regions. The energy—and therefore the mass—of a gravitational field is a slippery eel indeed, and refuses to be pinned down in any clear location. Nevertheless, it must be taken seriously. It is certainly *there*, and has to be taken into account in order that the concept of mass can be conserved overall. There is even a good (and positive) measure of mass (due to Bondi 1960, Bondi *et al.* 1962, and Sachs 1962) which applies to gravitational waves, but the non-locality is such that it turns out that this measure can be non-zero in regions where the space–time is actually completely free of curvature. Indeed, the region between two impulsive bursts of gravitational radiation can be completely flat, but the total energy carried by the wave as a whole simply depends on the extent of this flat region, i.e. on the distance between the two bursts. (See Penrose 1966 and Penrose and Rindler 1986; this last reference also describes a 'quasi-local' definition which does take into account the non-local properties of gravitational energy in finite regions.) It seems that, if this mass–energy is to be located at all, it must in such cases be in this flat empty space—a region completely free of matter or fields of any kind. Our 'quantity of matter' is now either there, in the emptiest of empty regions, or it is nowhere at all.

Thus, we see that even before we need consider the mysterious effects of quantum theory, our theories of physics tell us that there is something very odd and counter-intuitive about the nature of matter. We cannot at all draw a clear dividing line between what we call 'matter' or 'substance' and what we call 'empty space'—supposedly, the voids entirely free of matter of any kind. Matter and space are not totally separate types of entity. Actual substance need not be clearly localized in space. These are hints that our treasured intuitive views as to the nature of physical reality are less close to the truth than one would have thought. The nebulous and non-local additional features that the quantum theory injects into our picture of the world greatly strengthen this conclusion, but such conclusions must already be drawn on the basis of classical theory. We must expect, also, that future theory will provide us with yet further shocks to our cherished intuitions.

References

Bondi, H. (1960), 'Gravitational Waves in General Relativity', *Nature* London, no. 186: 535.

Bondi, H., van der Burg, M. G. J., and Metzner, A. W. K. (1962), ‘Gravitational Waves in General Relativity, VII: Waves from Axisymmetric Isolated Systems’, *Proc. Roy. Soc. London* A269: 21–52.

Einstein, A. (1905), *Annalen der Physik* 17: 639.

Maxwell, J. C. (1865), ‘A Dynamical Theory of the Electromagnetic Field’, *Phil. Trans. Roy. Soc.* 155: 459–512.

—— (1873), *A Treatise on Electricity and Magnetism*, Oxford University Press.

Penrose, R. (1966), ‘General–Relativistic Energy Flux and Elementary Optics’, in B. Hoffman (ed.), *Perspectives in Geometry and Relativity*, Indiana University Press, Bloomington, pp. 259–74.

—— and Rindler, W. (1986), *Spinors and Space-Time*, Vol. 2: *Spinor and Twistor Methods in Space-Time Geometry*, Cambridge University Press, p. 427.

Sachs, R. K. (1962), ‘Gravitational Waves in General Relativity, VIII: Waves in Asymptotically Flat Space-time’, *Proc. Roy. Soc. London* A270: 103–26.

205

TWISTORS AS SPIN 3/2 CHARGES CONTINUED: *SL*(3,*C*) BUNDLES

Twistor Newsletter **33** (1991). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

This is, in my view, a not fully convincing attempt to understand the puzzling issues thrown up by Chapters 202 and 203.

152 Twistors as spin 3/2 charges continued: *SL*(3,*C*) bundles. *Twistor Newsletter* **33** (1991). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

§III.4.22 Twistor as Spin $\frac{3}{2}$ Charges Continued: $SL(3, C)$ Bundles *by R. Penrose*

(𝕋ℕ 33, November 1991)

In §III.4.20 (and in Penrose 1992) I made the suggestion that a concept of twistor, appropriate for Ricci-flat space-times generally, would be as a (conserved) *charge* for massless helicity $\frac{3}{2}$ fields. This suggestion is based on the following two observations:

(1) $R_{ab} = 0$ provides the consistency condition (in an appropriate sense) for the existence and propagation of such fields in curved space-time $\mathscr{M}$ (Buchdahl 1958, Deser & Zumino 1976, Julia 1982);

Twistors as spin 3/2 charges continued: *SL*(3,*C*) bundles. *Twistor Newsletter* **33** (1991).
Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston,
P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

(2) In Minkowski space-time $\mathbb{M}$, the space of charges for such fields in naturally identified with the twistor space $\mathbb{T}$ of $\mathbb{M}$.

Thus, it would seem that if the appropriate concept of 'charge' for massless helicity $\frac{3}{2}$ fields in a Ricci-flat $\mathscr{M}$ could be found, then this should provide the long-sought concept of twistor appropriate to the vacuum Einstein equations. By analogy with the 'nonlinear graviton construction', the long-term programme is thus as follows:

(a) Find this concept of twistor for Ricci-flat $\mathscr{M}$;

(b) Find out how to characterize the geometry of this resulting twistor space $\mathscr{J}$ — the hope being that the general such $\mathscr{J}$ can be constructed with free functions;

(c) Find how to reconstruct $\mathscr{M}$ from $\mathscr{J}$.

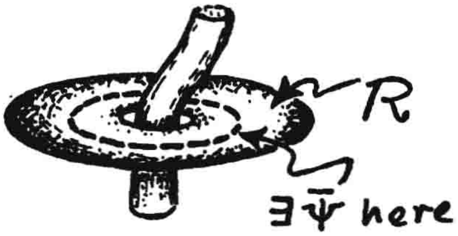
The concept of charge in (2) above is closely analogous to the concept of energy-momentum/angular momentum that arises for linear gravity — the sources for a spin-2 massless fields in $\mathbb{M}$ (conserved). If we think of this linear gravity spin-2 field as being described by a (Weyl tensor-like) object

$$K_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\Psi}_{A'B'C'D'},$$

where $\bar{\Psi}_{A'B'C'D'} = \bar{\Psi}_{(A'B'C'D')}$ satisfies the massless helicity 2 field equation

$$\nabla^{AA'}\bar{\Psi}_{A'B'C'D'} = 0$$

in some region $\mathscr{R}$ of $\mathbb{M}$ surrounding a world-tube



then we can obtain the 'charge' values for the tube by spin-lowering with a dual twistor twice ($2 \xrightarrow{\text{dual twistor}} \frac{3}{2} \xrightarrow{\text{dual twistor}} 1$) to obtain a self-dual Maxwell field, whose charges (electric + i × magnetic) can then be obtained by a Gauss integral, as is done in the quasi-local mass construction. Since two dual twistor are involved, the charges for linear gravity come out as the components of a symmetric valence 2 twistor $A^{\alpha\beta}$. The similar construction for a massless helicity $\frac{3}{2}$ field $\Psi_{A'B'C'}$ $(=\Psi_{(A'B'C')};\ \nabla^{AA'}\Psi_{A'B'C'}=0)$ involves only one dual twistor $(\frac{3}{2}\xrightarrow{\text{dual twistor}}1)$ so we get the charges as components of a valence 1 twistor, say Z^α.

For a general unrestricted $A^{\alpha\beta}(=A^{(\alpha\beta)})$ there would be three types of components, with dependence upon position x^a:

quadratic in x^a: angular momentum (6 real components) (primary part of $A^{\alpha\beta}$)
linear in x^a: 4-momentum (4 real components) and 'magnetic' (NUT) 4-momentum (4 real components)
independent of x^a: 'anti-angular momentum' (6 real components) (projection part of $A^{\alpha\beta}$)
Total: 20 real components (=10 complex components)

But when there is an appropriate potential h_{ab} for $\bar{\Psi}_{A'B'C'D'}$ in $\mathscr{R}$ (with

$$\bar{\Psi}_{A'B'C'D'}=\frac{1}{2}\nabla^{A}_{(A'}\nabla^{B}_{B'}h_{C'D')AB};$$

cf. Penrose and Rindler 1984, pp. 364), or, equivalently, when the sources in the world-tube come from a (linearized) energy-momentum tensor (cf. Penrose and Rindler 1986 §6.2), we have $A^{\alpha\beta}I_{\beta\gamma}=\bar{A}_{\gamma\beta}I^{\beta\alpha}$; then the 'anti-angular momentum' and 'magnetic' (NUT) components vanish and the order of dependence on position for the remaining components is reduced by one; linear in x^a for angular momentum; independent of x^a for 4-momentum.

All this is rather analogous to what occurs for spin $\frac{3}{2}$. Here we have

linear in x^a: ω-part of Z^α (primary part) 4 real components
independent of x^a: π-part of Z^α (projection part) 4 real components
Total: (8 real cpts.=) 4 complex components

Somewhat analogous to the condition $A^{\alpha\beta}I_{\beta\gamma}=\bar{A}_{\gamma\beta}I^{\beta\alpha}$, arising from the existence of h_{ab} globally in $\mathscr{R}$, is the condition $Z^\alpha I_{\alpha\beta}=0$ (i.e., $\pi_{A'}=0\ \therefore\ \omega^A=\text{const}$) arising from the existence of the first potential $\gamma^{A}_{B'C'}$ for $\Psi_{A'B'C'}$ globally throughout $\mathscr{R}$. Here we can adopt either the Dirac form $\boxed{D}$ of the potential, for which the symmetry

$$\boxed{D}\qquad\qquad \gamma^{A}_{B'C'}=\gamma^{A}_{C'B'}$$

holds, or else the Rarita-Schwinger form $\boxed{\text{R-S}}$, for which this symmetry is not imposed. The equations to be satisfied are

$$\boxed{D}\qquad\qquad \nabla^{BB'}\gamma^{A}_{B'C'}=0$$

Twistors as spin 3/2 charges continued: *SL*(3,*C*) bundles. *Twistor Newsletter* **33** (1991).
Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston,
P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

or

$$\boxed{\text{R-S}} \qquad \begin{cases} \nabla^{B'(B}\gamma^{A)}_{B'C'} = 0 & \cdots\cdots \quad (1)\\ \varepsilon^{B'C'}\nabla_{A(A'}\gamma^{A}_{B')C'} = 0 & \cdots\cdots \quad (2)\end{cases}$$

with gauge freedom $\gamma^{A}_{B'C'} \mapsto \gamma^{A}_{B'C'} + \nabla^{A}_{B'}\nu_{C'}$, where

$$\boxed{D} \qquad \nabla^{AA'}\nu_{A'} = 0$$

(but $\nu_{A'}$ is *free* in the $\boxed{\text{R-S}}$ case). These equations also work in *Ricci-flat* $\mathscr{M}$.

We can also define the 'field' $\Psi_{A'B'C'}$ $(= \Psi_{(A'B'C')})$ by

$$\boxed{D} \qquad \Psi_{A'B'C'} = \nabla_{AA'}\gamma^{A}_{B'C'}$$

or

$$\boxed{\text{R-S}} \qquad \Psi_{A'B'C'} = \nabla_{A(A'}\gamma^{A}_{B')C'}$$

In $\mathbb{M}$, $\Psi_{A'B'C'}$ is also gauge invariant and satisfies the field equation $\nabla^{AA'}\Psi_{A'B'C'}$. However, for this to be true in Ricci-flat $\mathscr{M}$, we also need the condition that the Weyl tensor is anti-self-dual (A.S.D.; i.e., $\bar{\Psi}_{A'B'C'D'} = 0$). In that case, we define the 'π-space' for twistors by

$$\mathbb{S}_{A'} = \{\text{glob.}\Psi\text{s}\}/\{\text{glob.}\gamma\text{s}\}$$

where '$\{\text{glob.}\Psi\text{s}\}$' means the space of fields $\Psi_{A'B'C'}$ that are global throughout $\mathscr{R}$, and correspondingly for '$\{\text{glob.}\gamma\text{s}\}$'. This does not work when the Weyl tensor is not A.S.D.; however, even though there is apparently a natural-looking suggestion for a meaning for '$\{\text{glob.}\Psi\text{s}\}$' in terms of a patchwork of gauge-equivalent γs (see Penrose 1992). Taking an open covering $\{\mathscr{U}\}$ for $\mathscr{R}$ we can try to define $\overset{i}{\gamma}$ on $\mathscr{U}$ where $\overset{i}{\gamma} - \overset{j}{\gamma}$ is pure gauge on each $\mathscr{U} \cap \mathscr{U}$. Unfortunately we find, in general, that any such 'global ψ' is equivalent to a global γ.

It is instructive to compare this with the spin-2 case where we can consider choosing $\overset{i}{h}_{ab}$ on $\mathscr{U}$, to represent a linearized perturbation of the matrix, with

$$\overset{i}{h}_{ab} - \overset{j}{h}_{ab} = 2\nabla_{(a}\xi_{b)} \quad \text{on} \quad \mathscr{U} \cap \mathscr{U}$$

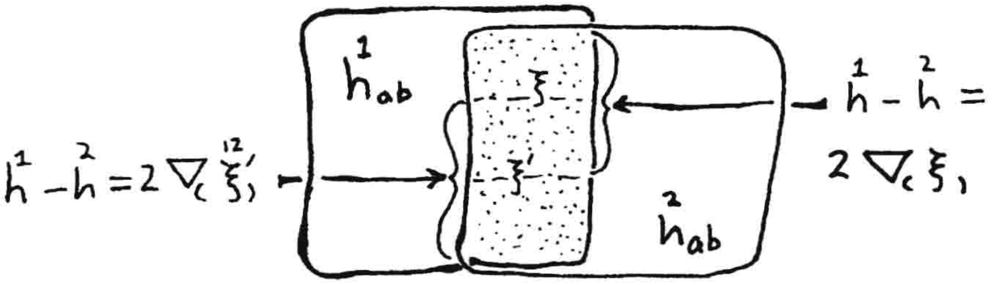
(gauge equivalence). We find that for this to represent a genuine metric perturbation on the whole of $\mathscr{R}$, we must have, for every triple overlap

$$\overset{ij}{\xi} + \overset{jk}{\xi} + \overset{ki}{\xi} = 0 \quad \text{on} \quad \mathscr{U} \cap \mathscr{U} \cap \mathscr{U}$$

(and also, in accordance with the phenomenon for γs, as described in §III.4.20, does *not* count as 'gauge equivalence' between $\overset{1}{h}$ and $\overset{2}{h}$ unless $\overset{12}{\xi} = \overset{12}{\xi}{}'$ in the central region of the overlap in the figure below).

156 Twistors as spin 3/2 charges continued: *SL*(3,*C*) bundles. *Twistor Newsletter* **33** (1991).

Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001


$\overset{1}{h}_{ab}$
ξ
ξ'
$\overset{2}{h}_{ab}$
$\overset{1}{h}-\overset{2}{h}=2\nabla_{(c}\overset{12}{\xi'}{}_{)}$
$\overset{1}{h}-\overset{2}{h}=2\nabla_{(c}\xi_{)}$

This condition on triple overlaps serves to exclude 'magnetic mass' (i.e., a NUT-type source). It is analogous to ruling out '$\pi_{A'}$-charge' in the spin $\frac{3}{2}$ case, which is just what we do *not* want to rule out. This suggests that we do something in the spin $\frac{3}{2}$ case that is analogous to Misner's solution to the problem of eliminating the NUT ('Dirac-type') singular wires — the triple regions on which $\overset{12}{\xi}+\overset{23}{\xi}+\overset{31}{\xi}=0$ fails. Misner noted that NUT space could be made non-singular ('wire-free') by means of identifications in the time direction that introduced closed time-like curves in the space-time. This is only possible because the NUT space-time has a time-like Killing vector, and the symmetry in that direction allows identifications to be made.

If we are to try something similar in the general spin $\frac{3}{2}$ case, we would want to erect some kind of bundle over $\mathscr{M}$ which will admit the necessary extra symmetries in the fibre direction. We also need γ to have some kind of nonlinear interpretation (since we do not expect a linear π-space in the general case). The best bet appears to be to interpret γ as providing a *bundle connection.*

Take the fibre coordinates to be given by a spinor $\eta_{A'}$ and a scalar ξ. For a given small ϵ, we extend the ordinary (Christoffel-Levi-Civitá) covariant derivative ∇ on $\mathscr{M}$ to bundle-valued quantities according to

$$\nabla_{PP'}\begin{pmatrix}\eta_{A'}\\ \zeta\end{pmatrix}=\begin{pmatrix}\nabla_{PP'}\eta_{A'}\\ \nabla_{PP'}\zeta\end{pmatrix}-\epsilon\begin{pmatrix}0 & \gamma_{PP'A'}\\ \gamma^{B'}_{PP'} & 0\end{pmatrix}\begin{pmatrix}\eta_{B'}\\ \zeta\end{pmatrix},$$

the gauge transformations being

$$\begin{pmatrix}\eta_{A'}\\ \zeta\end{pmatrix}\mapsto\begin{pmatrix}\eta_{A'}\\ \zeta\end{pmatrix}+\epsilon\begin{pmatrix}0 & \nu_{A'}\\ \nu^{B'} & 0\end{pmatrix}\begin{pmatrix}\eta_{B'}\\ \zeta\end{pmatrix},$$

where terms of order ϵ^2 are being neglected. (There appears to be some relation to supersymmetry operations here, but as far as I can make out, what I am doing is not the same; also, an interesting limiting case of the above (studied by Paul Tod) occurs when one of the two off-diagonal entries in the matrices is set to zero, but this seems not to lead to the needed nonlinearities.) As they stand, the gauge transformations do not close under commutation, and for a consistent theory valid to all orders in ϵ we need to generalize to SL(3,$\mathbb{C}$) matrices before the commutation closes. Writing $\textcircled{A}$ for the three-dimensional indices occurring here, so $\eta_{\textcircled{A}} = \begin{pmatrix} \eta_{A'} \\ \zeta \end{pmatrix}$, etc., we have a connection defined according to:

$$\nabla_{PP'}\eta_{\textcircled{A}} = \begin{pmatrix} \nabla_{PP'}\eta_{A'} \\ \nabla_{PP'}\zeta \end{pmatrix} - \gamma_{PP'\textcircled{A}}{}^{\textcircled{B}}\eta_{\textcircled{B}}$$

with gauge transformations

$$\eta_{\textcircled{A}} \mapsto \eta_{\textcircled{A}} + \nu_{\textcircled{A}}{}^{\textcircled{B}}\eta_{\textcircled{B}}.$$

The $\nu_{\textcircled{A}}{}^{\textcircled{B}}$ are $SL(3,\mathbb{C})$-valued fields on $\mathscr{M}$, so

$$\epsilon^{\textcircled{P}\textcircled{Q}\textcircled{R}}\,\nu_{\textcircled{P}}{}^{\textcircled{A}}\,\nu_{\textcircled{Q}}{}^{\textcircled{B}}\,\nu_{\textcircled{R}}{}^{\textcircled{C}} = \epsilon^{\textcircled{A}\textcircled{B}\textcircled{C}},$$

where $\epsilon^{\textcircled{P}\textcircled{Q}\textcircled{R}}$ (and $\epsilon_{\textcircled{P}\textcircled{Q}\textcircled{R}}$) are Levi-Civitá objects, so

$$\epsilon^{\textcircled{A}\textcircled{B}\textcircled{C}}\,\overset{1}{\eta}_{\textcircled{A}}\,\overset{2}{\eta}_{\textcircled{B}}\,\overset{3}{\eta}_{\textcircled{C}} = \overset{1}{\eta}_{A'}\,\overset{2}{\eta}{}^{A'}\,\overset{3}{\zeta} + \overset{2}{\eta}_{A'}\,\overset{3}{\eta}{}^{A'}\,\overset{1}{\zeta} + \overset{3}{\eta}_{A'}\,\overset{1}{\eta}{}^{A'}\,\overset{2}{\zeta}.$$

The γs are likewise generalized, and we have an object like

$$\gamma_{PP'\textcircled{A}}{}^{\textcircled{B}} = \begin{pmatrix} \alpha_{PP'A'}{}^{B'} & \beta_{PP'A'} \\ \gamma_{PP'}{}^{B'} & \delta_{PP'} \end{pmatrix}.$$

The curvature is

$$K_{pq\textcircled{A}}{}^{\textcircled{B}} = 2\nabla_{[p}\gamma_{q]\textcircled{A}}{}^{\textcircled{B}} + 2\gamma_{[p|\textcircled{A}|}{}^{\textcircled{C}}\gamma_{q]\textcircled{C}}{}^{\textcircled{B}}.$$

It seems that we must consider this as a generalization of the $\boxed{\text{R-S}}$ rather than the $\boxed{D}$ form, since the commutators of $\boxed{D}$ gauge quantities cease to satisfy $\nabla^{A'}_{A}\nu_{A'} = 0$ or any other sensible local equation. One is led to speculate that an appropriate nonlinear generalization of the $\boxed{\text{R-S}}$ equations might well be

$${}^{-}K_{PQ\textcircled{A}}{}^{\textcircled{B}} = 0, \qquad {}^{+}K_{P'Q'\textcircled{A}}{}^{\textcircled{B}}\,\epsilon^{P'\textcircled{A}\textcircled{C}}\,\epsilon_{Q'\textcircled{B}\textcircled{D}} = 0$$

(or some such), where

$$K_{pq\textcircled{A}}{}^{\textcircled{B}} = {}^{-}K_{PQ\textcircled{A}}{}^{\textcircled{B}}\,\epsilon_{P'Q'} + \epsilon_{PQ}\,{}^{+}K_{P'Q'\textcircled{A}}{}^{\textcircled{B}}$$

in the splitting of the curvature into anti-self-dual and self-dual pieces, respectively. (These equations are respective analogues of 1 and 2.) Here,

$$\epsilon^{P'\textcircled{A}\textcircled{C}} = \epsilon^{\textcircled{P}\textcircled{A}\textcircled{C}}\, e_{\textcircled{P}}{}^{P'}, \qquad \epsilon_{Q'\textcircled{B}\textcircled{D}} = \epsilon_{\textcircled{Q}\textcircled{B}\textcircled{D}}\, e^{\textcircled{Q}}_{Q'},$$

where $e_{\textcircled{P}}{}^{P'}$ and $e_{Q'}{}^{\textcircled{Q}}$ are appropriate (degenerate) 'soldering' quantities relating the bundle directions with tangent (spinor) directions in $\mathscr{M}$, and there are presumably also equations on these, like perhaps $\nabla_{PP'}e_{\textcircled{P}}{}^{P} = 0$. This is very much 'work in progress', and the appropriate geometrical viewpoint has not fully come to light. (There are some possible relations between this type of bundle and other constructions of relevance to twistor theory. If these become clearer they will be reported on later.)

Note that the equation on $^{-}K^{\cdot}_{\ldots 0}{}^{0}$ tells us that we have a *self-dual connection* of a particular type. Thus, in the special case when $\mathscr{M}$ has a self-dual Weyl curvature, the Ward construction should lead to an interpretation of this connection in terms of $\mathscr{M}$'s *dual* twistor space, as if the charges could be defined for this connection, we have a different angle on the googly.

I am grateful particularly to Abhay Ashtekar, Lee Smolin, Ted Newman, George Sparling and Paul Tod for discussions.

References

Buchdahl, H.A. (1958) On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field, Nuovo Cimento, **10**, 335–7.

Deser, S. & Zumino, B. (1976) Consistent supergravity, Phys. Lett., **62B**, no. 4, 335–7.

Julia, B. (1982) Linear System associated with Einstein's equations, Comptes Rendus Acad. Sci. Paris, Sér. **II**, no. 2, 113–6.

Penrose, R. (1992) in *Gravitation and Modern Cosmology*, eds., A. Zichichi, N. de Sabbata, and N. Sánchez, Plenum Press, New York.

Penrose, R. & Rindler, W. (1984, 1986) *Spinors and Space-Time*, Vols. 1, 2, Cambridge University Press, Cambridge.

206

RESPONSE TO TONY DODD'S 'GODEL, PENROSE AND THE POSSIBILITY OF AI'

Artificial Intelligence Review **5** (1991), no. 4, 235

Without access to Dodd's article, it may be a little difficult to fathom the points at issue completely here, but Dodd's original article can be found on pp. 187–199 of the above publication.

Response Time

Response to Tony Dodd's Gödel, Penrose and the Possibility of AI

Roger Penrose
Mathematical Institute, University of Oxford

Tony Dodd (*Artificial Intelligence Review* **5.3**) provides a fine exposition of the issues raised by Godel's theorem and I am in full agreement with virtually all that he says — except for the final conclusion!

A minor point of disagreement is that one *can* actually formulate all that is needed entirely in terms of algorithms rather than formal systems (despite what Dodd says), including the necessary diagonalization (see the section on 'How to outdo an algorithm' in Chapter 2 of *The Emperor's New Mind* — but I intend to write a more complete account of this at a later date). I agree with Dodd that my argument *does* immediately imply that 'mathematics is not an axiomatic discipline', but unlike him, I have no trouble in accepting this 'dramatic conclusion'. Axioms and formal systems are useful in clarifying and systematizing broad areas of mathematical reasoning, but they do not themselves constitute the ultimate basis of mathematical belief. I had thought that this was indeed the main thrust of Gödel's theorem. I am afraid that, unlike Dodd, I *do* find it 'absurd to use a theory without believing its consistency' if by using a theory one is implicitly believing it to be *sound*, e.g. that its implications with regard to number theory are actually true — and, in particular, that it does not provide us with a 'proof' that 2 = 3. If we believe that a formal system F does not prove 2 = 3, then we must believe that F is, in fact, *consistent* so, by Godel's theorem, that particular F cannot be a complete formalization our mathematical insights — *whatever* F might be.

An 'unpleasant consequence' is indeed that 'the correctness of a mathematical proof cannot be checked mechanically' if by a proof we mean an argument that may be partly of an informal nature such as the 'Gödelian' type of reasoning just referred to. Dodd quite correctly emphasizes that the difference between proof-finding and proof-checking is merely a matter of complexity and not computability. As I see it, the Gödel argument does seem to show us that, at root, our mathematical beliefs depend upon some kind of 'Platonic intuition' but these beliefs are continually being sharpened, strengthened and corrected by formal reasoning. Mathematicians can certainly be mistaken at times, but mistakes can be rooted out and ultimately recognized as such. Dodd does not seem to like this general Platonic conclusion and his escape is 'that the strength of our mathematical beliefs shades off by degrees'. This is perhaps a possible way out, but I do not feel that it answers Godel's conundrum, as it is surely *obvious* that belief of soundness implies a belief in consistency. There does not seem to be much shading off in one's beliefs in passing from a belief in soundness to a belief in consistency.

207

THE BIGGEST ENIGMA

Chemtech **22** (1992), no. 1, 28–31

Book reviews have, as a general rule, not been included in these Collected Works, but this one, from the *New York Review of Books*, is more of a scientific discourse on the nature of quantum mechanics, and seems worthy of inclusion.

The biggest enigma

Roger Penrose

Of all the remarkable advances in scientific understanding that have been achieved over many centuries, the theory of quantum mechanics has been the most enigmatic. Introduced in the first quarter of the twentieth century, the theory grew to a beautiful mathematical structure that became the most revolutionary, broad-ranging, and successful of modern theories. Yet despite the multitude of insights that quantum mechanics has given us, it remains the most baffling of all successful theories. Quantum mechanics has such mathematical precision and range of predictive power that it provides explanations for the stability of atoms and solids; for phase changes such as melting, freezing, and boiling; for the colors of heated materials, including a detailed description of the puzzling phenomenon of spectral lines; for low-temperature superconductivity and superfluidity; and for the behavior of lasers, transistors, and television screens, not to mention the whole of chemistry and much of biology and genetics. How can it be that such a successful theory can still remain an enigma, even to the greatest of physical scientists?

It may be recalled that, for a time, that other great revolutionary physical theory, Einstein's general relativity, introduced in about 1915, was supposed to have been understood by but a handful of people; yet today that theory is not considered to be exceptionally hard to grasp in its entirety, and it provides us with a mathematically consistent and philosophically satisfying picture of the universe. We may ask why the same has not happened with quantum mechanics—for indeed it has not: Even one of the greatest quantum physicists of modern times, Richard Feynman, was led to write a few years ago in his book on quantum electrodynamics for the lay person, "I don't understand it, nobody does" (*1*). Yet the answers that different physicists might give to this question would vary greatly, depending upon the views they might hold on the quantum theory.

It is not that the theory is particularly difficult to comprehend as a mathematical structure. Rather, the theory exhibits a baffling mixture of extreme physical precision with utter absurdity, with regard to the way it relates to the macroscopic world that we actually perceive. The irony of the situation was once captured succinctly by Robert Wald, a colleague from the University of Chicago, in a dinner-table discussion about attitudes to the validity of the theory: "If you really believe in quantum mechanics then you can't take it seriously." This viewpoint expresses essentially that of those who support the "conventional wisdom" referred to as the "Copenhagen interpretation." They indeed claim that the very precise formalism of the theory is not to be taken seriously as a picture of actual "reality." They often assert, accordingly, that the whole question of quantum reality is a nonquestion. One should not think of the theory as providing us with a picture of actuality, they argue, but merely as giving us a calculational procedure that accurately provides the correct mathematical probabilities for the different possible outcomes of experiments. This, they say, is all that we should ask of a theory, and we need not ask questions about "reality." We do not need an understanding of the "actual" nature of the world; it is amply sufficient for our theory to make accurate "predictions"—something that quantum mechanics is indeed supremely good at.

Of course there is a macroscopic world that we can directly perceive, and that must be assigned some "reality," but the submicroscopic world of quantum phenomena need not be. This was the position of Niels Bohr, originator in 1913 of the quantum model of the atom and of the "old" quantum theory that preceded the more refined theory of Heisenberg and Schrödinger, which came about in 1925–1926. Bohr also thought deeply about the newer quantum ideas and became the

Niels Bohr

Albert Einstein

Drawings by David Levine. Reprinted with permission from *The New York Review of Books* Copyright 1991 Nyrev. Inc.

"high priest" of the Copenhagen viewpoint. Among other things, he was insistent that one should assign no "reality" to the quantum description of a state of a system between the "measurements" made by scientists, the results of these measurements themselves being the only things one should attach reality to.

Such a defeatist view was not shared by all originators of quantum theory. In particular, Albert Einstein—who had himself been responsible for the first ideas leading ultimately to the quantum realization that particles were waves and waves were particles—found such a physical standpoint abhorrent (as did Schrödinger). To Einstein, the fundamental prerequisite of a scientific view of the world was a belief in a reality "out there" whose nature, no matter how strange, must be independent of how we might choose to examine it—even though by examining it we might indeed change that reality. This "objective reality" must hold at the quantum level of things—the submicroscopic world of molecules, atoms, and nuclear particles—just as it does at the macroscopic level of everyday experience. The aim of fundamental science, according to Einstein, is to find out what the world is actually like, and to discover the laws according to which that world behaves at all levels, including the level of quantum phenomena. If quantum theory does not provide a realistic view of events at the quantum level, the theory must be incomplete or perhaps even just plain wrong. Thus Einstein was led to reject the quantum theory as not providing the appropriate description at submicroscopic scales.

Einstein's objections to the indeterminism in quantum theory are well known, and he wrote to Max Born, in 1926: "The theory produces a good deal but hardly brings us closer to the secret of the Old One. I am at all events convinced that He does not play dice." Yet this indeterminism did not appear to have been Einstein's main objection to the theory. It was the apparent subjectivity of "physical reality," inherent in the Copenhagen interpretation, that he could not abide. Einstein made many attempts to show that quantum theory was actually inconsistent, and up to and around 1930 he was involved in many arguments with Niels Bohr.

Although the consensus was that Bohr won these arguments, Einstein never gave up. In 1935 he produced, with Boris Podolsky and Nathan Rosen (2), a paper pointing out a disturbing implication of quantum mechanics (now called the EPR effect after the three men): Certain experimental situations could be set up in which the close choice of the way in which a particle is to be observed in one region of space would seemingly instantaneously affect the state of a second particle that could be very distant from the first. In fact the effect would be totally independent of the distance between the two particles.

This was disturbing for Einstein, and he referred to the effect as "spooky action at a distance." Not only did this "action" not get smaller as the distance increased, as would a normal, self-respecting force, but the instantaneous nature of this action seemed incompatible with the picture of the world that Einstein himself had provided in his special theory of relativity. In fact there is no implied "observational action" in Einstein's special theory, but there is a serious conceptual conflict.

Einstein's followers, most notably David Bohm in 1952 (*3*) (and also Prince Louis Victor de Broglie [*4*], the founder of the idea of a universal wave/particle duality), later developed a theory of "hidden variables" according to which the uncertainties in quantum behavior could arise because of the random motions of some underlying substructure or "parts" that cannot be directly observed but provide the needed "reality" for subquantum level objects. To be consistent with the effect described above, however, it would be necessary for any "realistic" view of quantum mechanics, such as the de Broglie–Bohm theory, to be essentially nonlocal. Thus the "spooky" instantaneous actions would have to be built in as part of "reality" according to the theory. This would be very much in conflict with the spirit of Einstein's relativity theory.

For a while, physicists did not worry themselves much with such matters, these being regarded more as questions of philosophy than of serious physics. Things began to change, however, when in 1964 the Northern Irish physicist John Bell published a paper pointing out a mathematical inequality that showed that the "spooky action" of quantum theory could actually be put to experimental test (*5*). A version of the EPR effect put forward by Bohm was subsequently used as the basis of a number of experiments, most notably one performed by Alain Aspect (*6*) in Paris in the early 1980s which convincingly confirmed that this "spooky" nonlocal quantum effect is, in a queer sense, really true, and is now part of observed physical reality.

Though there are probably more different attitudes to quantum theory than there are quantum physicists, the particular polarity of views expressed by Einstein and Bohr encompasses an important part of this spectrum of opinion. Jeremy Bernstein, in his new book *Quantum Profiles* (*7*), provides us with an unusual insight into this controversy by addressing the physical issues indirectly, through the opinions and attitudes of some of the major personalities involved. There are just three chapters (two of which have epilogues), one on Bell, one on John Wheeler (an influential disciple of Bohr's), and one on Einstein's lifelong friend and colleague Michele Besso. The chapters on Bell and Wheeler are based mainly on personal interviews, while the one on Besso is based on his extensive correspondence with Einstein.

I feel that it is a matter of tremendous good fortune that Bernstein was able to interview John Bell when he did—in January 1989—for Bell is no longer with us, having unexpectedly died of a cerebral hemorrhage in October 1990, at the age of 62. This is a tragic loss. In my own opinion, Bell's contributions to our understanding of quantum mechanics have been unique and fundamental. He opened the way to a striking class of new experiments which, in the words of the physicist and philosopher Abner Shimony, provide, perhaps for the first time, experiments that can actually test one's philosophical viewpoint concerning the nature of reality. In my opinion, there is an additional tragedy in Bell's loss. He could not but be respected by all who thought deeply about quantum mechanics, whatever their persuasions. Even the "Copenhagen hard-liners" had to respect his views. In the later years of his life, Bell gave strong support and encouragement to the ideas of some of those who, in recent years, have dared to challenge conventional quantum theory with model rival theories of their own. Although all of these new theories have their weaknesses, Bell has made this activity respectable. I believe that this is vitally important. For, before too long, we shall indeed need a new theory. I profoundly hope that the new-found respectability for such searchings will not die with Bell. In his account, Bernstein gives us a warm and friendly insight into the personality of this subtly humorous, gentle, and admirable man, in addition to letting us in on Bell's profound and clear views on the puzzles of quantum mechanics and on Bell's insightful commentaries concerning Bohr's and Einstein's views.

The chapter on John Wheeler is also illuminating. Wheeler is a fascinating man: a paradoxical union within one personality of the revolutionary and the profoundly conservative—a perfect gentleman with a wild delight in explosives (and a leading figure in the development of the hydrogen bomb). I liked Wheeler's summary of John Bell's approach to quantum theory: that Bell would rather be clear and wrong than foggy and right. There is certainly something of the Einstein/Bell type of view in the first of these alternatives, as opposed to the Bohr/Wheeler type of view in the second. Bell always made it clear that his sympathies lay with Einstein in his controversy with Bohr, and that Einstein's philosophical standpoint was far superior to that of Bohr—even though it was Bell's own insights that finally led to a refutation of Einstein's "local realism." It just seemed to Bell that it was a "pity" that Bohr has in this important respect, at least, turned out to be right!

There is indeed a profound fogginess in Bohr's and

Wheeler's expression of views. For myself, I do not find this obscurity helpful, and in this respect my sympathies are definitely with Bell and Einstein. Yet there is a deep point, also, in Wheeler's stressing the mathematical elegance of the formalism of quantum theory. Although it is at odds with good philosophy, quantum theory enjoys a profound harmony with mathematics.

My personal belief is that the resolution of these puzzles can never be achieved within the theory as it stands. We must search for a new theory with a mathematical elegance even greater than that of present-day quantum mechanics—to which quantum theory can be only an excellent approximation—and which will also make profound philosophical sense. I am sure that Bell would agree with this. I am not so sure about Wheeler, who seems more concerned with seeing "why" quantum mechanics is "utterly right" than in exploring how it might actually turn out eventually to be wrong.

In the chapter on Besso, we learn more about Einstein himself, his generous but dedicated personality, and his powerful views about quantum mechanics. There is something here of historical interest both in relation to Einstein's scientific persona and the strains of his family life. It is a useful supplement to the existing biographies of Einstein.

I found Bernstein's book fascinating to read (perhaps because I have known the remarkable scientists featured in two of the three chapters). Perhaps I learned less that was new about their views on quantum mechanics than I had hoped I might, but I did gain some new insights. However, there are several curious errors or misprints throughout the book that caught my attention, such as the conflation "Max Bohr," and there are some others that are also genuinely confusing. Also, in his description of Galileo's Dialogues, Bernstein asserts that it is "Sagredo" who puts forward Galileo's own views, whereas I had always thought this was normally "Salviati." Bernstein's book is valuable in that it emphasizes the fact that quantum theory is science and not mysticism, and that one should be extremely wary of the frequent associations that are made between these two very different ways of thinking about the world. The book provides a valuable addition to an important debate about our understanding of the universe at its most fundamental level.

References

(1) Feynman, R. *QED, The Strange Theory of Light and Matter*; Princeton University Press: Princeton, N. J., 1985, p. 9.
(2) Einstein, A.; Podolsky, P.; Rosen, N. In *Quantum Theory and Measurement*; Wheeler, J. A.; Zurek, W. H., Eds. Princeton University Press: Princeton, N. J., 1983. Originally in *Phys. Rev.* **1935**, *47*, 454.
(3) Bohm, D. In *Quantum Theory and Measurement*, Wheeler, J. A.; Zurek, W. H., Eds. Princeton University Press: Princeton, N. J., 1983. Originally in *Phys. Rev.* **1952**, *85*, 166.
(4) de Broglie, L. *Tentative d'interpretation causale et nonlineaire de la mechnique ondulatoire;* Gauthier-Villars: Paris, 1956.
(5) Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics*; Cambridge University Press: Cambridge, 1987.
(6) Aspect, A.; Grangier, P. In *Quantum Concepts in Space and Time*; Penrose, R.; Isham, C. Oxford University Press: Oxford, 1986.
(7) Bernstein, J. *Quantum Profiles*; Princeton University Press: Princeton, N. J., 1990, 178 pp. $17.95.

Reprinted with permission from *The New York Review of Books*.
Copyright 1991 Nyrev, Inc.

Roger Penrose is the Rouse Ball Professor of Mathematics at Oxford University (Mathematical Institute, 24-29 St. Giles, Oxford OX1 3LB, England; 0865-273546). He has received a number of prizes and awards, including the 1988 Wolf Prize which he shared with Stephen Hawking. He has attacked the concept of strong artificial intelligence in his book *The Emperor's New Mind* (Oxford University Press, 1989) on the basis of quantum mechanics.

208

$\mathcal{H}$-SPACE AND TWISTORS

Recent advances in general relativity: Essays in honor of Ted Newman (1990), 6–25, Einstein Stud., **4**, Birkhauser Boston, 1992

This provides something of a historical account of the birth of $\mathcal{H}$-space and of its role in the discovery of the 'non-linear graviton'.

$\mathcal{H}$-space and twistors. *Recent advances in general relativity: Essays in honor of Ted Newman* (1990), 6–25, Einstein Stud., **4**, 1992

$\mathcal{H}$-Space and Twistors

Roger Penrose

It is a particular pleasure for me to have this opportunity to express with this article some of the heartfelt admiration that I have for Ted Newman. As well as being a very fine and inspirational scientist and a warm and stimulating human being, he is the person from whom I have gained the most in long-term scientific collaboration, certainly in the sense of the relative number of articles that he has coauthored with me (although I appreciate that the converse of this, in Ted's case, is far from true). I have no intention of attempting to summarize the work that we have accomplished together, nor to list the numerous ways in which Ted has been a continuing and lasting inspiration to me. Instead, I shall concentrate on just one of the most important of those inspirational insights that Ted has given us, and I shall try to indicate something of the deep implications that this insight has provided.

1. The Bondi–Metzner–Sachs Group and the Structure of $\mathscr{I}^+$

I shall start by describing a little of the background to Ted's seminal concept (Newman 1976) of $\mathcal{H}$-space (cf. Ko, Ludvigsen, Newman, and Tod 1981 for a comprehensive study of $\mathcal{H}$-space). In around 1960, Hermann Bondi, in conjunction with others, notably Ray Sachs (Bondi 1960; Bondi, van der Burg, and Metzner 1962; Sachs 1962a), succeeded in providing the physically appropriate definition for the total mass (and implicitly 4-momentum; Sachs 1962b; cf. also Penrose 1963) of a gravitationally radiating isolated physical system in which the separate, necessarily positive, contribution from the radiation itself could be identified. This depended upon the existence of an unambiguously defined *translation group* that had to be singled out from the entire group of asymptotic symmetries of the space-time $\mathcal{M}$ (see Sachs 1962b), where the entire asymptotic symmetry group, which has come to be known as the *Bondi–Metzner–Sachs (BMS) group*, can be identified as the basic symmetry group of *future null infinity* $\mathscr{I}^+$(Penrose 1963, 1974).

As impressive as this achievement was, it left unanswered the corresponding problem of finding the appropriate definition of *angular* momentum. An essential difficulty has its roots in the problem of finding the *full* (restricted) Poincaré group—rather than just the translation group—as a natural subgroup of the BMS group. $\mathcal{H}$-space arose out of Ted's determined efforts to circumvent this problem. Although these efforts were not rewarded by a successful solution of the angular momentum problem, they led, via $\mathcal{H}$-space, to something even more valuable.

It will be helpful to review the basics of the geometry of $\mathscr{I}^+$ and to remind ourselves of the way that the BMS group arises from it. (For further details, see Penrose and Rindler 1986, Section 9.8.) The topology of $\mathscr{I}^+$ in the standard asymptotically flat situation is $S^2 \times \mathbb{R}$, and we can take the various "copies" of $\mathbb{R}$ in this product to be the null geodesics (generators) of $\mathscr{I}^+$. It is convenient to choose the conformal factor for $\mathscr{I}^+$ in such a way that the generators for $\mathscr{I}^+$ are parallel (i.e., with $\rho' = 0$, where we also have $\sigma' = 0$). Then all cross sections of $\mathscr{I}^+$—or *cuts*, in Ted's descriptive terminology—have the same metric, and we can choose this metric to be that of a unit 2-sphere. The BMS group consists of self-transformations of $\mathscr{I}^+$ that preserve its (appropriately defined) conformal structure. BMS transformations involve conformal motions of this S^2; those that are the identity on S^2 are the *supertranslations* of $\mathscr{I}^+$, where the *translations* are singled out as a special 4-parameter subsystem of supertranslations. We can interpret the problem of isolating an appropriate (restricted) Poincaré subgroup of the BMS group as that of determining which BMS transformations are to be considered to be "supertranslation free." We can rephrase this problem as the singling out of a specific 4-parameter family of cuts; the *good cuts* (in Ted's terminology). The idea would be that the required supertranslation-free subgroup would be the one preserving this family of good cuts. The model providing the grounds for this expectation is Minkowski space $\mathbf{M}$. Here, the good cuts of $\mathscr{I}^+$ arise as the intersections of $\mathscr{I}^+$ with the future light cones of the points of $\mathbf{M}$. Since the Poincaré group sends points of $\mathbf{M}$ to other points of $\mathbf{M}$, the corresponding subgroup of $\mathscr{I}^+$ sends good cuts to good cuts.

How are we to define the good cuts for a general radiating asymptotically flat space-time $\mathcal{M}$? To choose the intersections with $\mathscr{I}^+$ of the future light cones of points in $\mathcal{M}$ would not be appropriate for a variety of reasons. In the first place, these intersections would frequently not be cuts of $\mathcal{I}^+$ at all because of the presence of caustics and crossing regions in these cones. (This would happen even for the space-time representing a single isolated body, for example, a star, since the light cones of points farther away than the focal length of the star's gravitational field—according to its lensing effect—would encounter such caustics. For the field of *two* stars of sufficient separation, there would be no points whose light cones fail to encounter such caustics.) But, the situation is, in a sense, worse than this, because even when we do get actual cuts (cross sections) of $\mathcal{I}^+$, there will in general be no BMS symmetries that take the cuts into one another. (There will be supertranslations taking particular cuts into other particular cuts, but the same particular supertranslations would not work for other pairs of cuts.)

One possible way out of this impasse is not to talk about good cuts generally, but only at the remote past, or else the remote future, on $\mathcal{I}^+$ (i.e., near i^0 or else near i^+, respectively). This approach was explored by Ted and me in our 1966 article (Newman and Penrose 1966), in which we also introduced ð ("eth"). We argued, on the basis of linearized gravity theory, that it is physically reasonable to assume that the appropriate curvature quantities fall off fast enough toward i^0 (or toward i^+) on $\mathcal{I}^+$ for there to be a 4-parameter family of good cuts near i^0 (or i^+), where "near" means in the limit as i^0 (or i^+) is approached along $\mathcal{I}^+$ and where we use the term "good cut" in the sense that the *shear* σ^0 vanishes on the cut. Here the quantity σ^0 refers to the asymptotic shear of the null hypersurface that meets $\mathcal{I}^+$ locally in the cut.

This procedure does work in the sense that a natural (restricted) Poincaré subgroup of the BMS group can be singled out as the subgroup preserving these asymptotic good cuts, but it is not altogether satisfactory physically. There is, of course, the awkwardness of getting *two*, almost always different, Poincaré subgroups of the BMS group, depending on whether we use i^0 or i^+ as the place where we examine the cuts. But, the problems from the physical point of view are really much worse than this. Suppose, for example, that we have an initially stationary isolated system that emits a burst of gravitational radiation, then returns closely to stationarity, radiates again, and finally settles down closely to a stationary state once more (Figure 1). One would like at least to be able to assign (perhaps only approximately) an angular momentum to the system while it is in its intermediate closely stationary state. But, neither of the two Poincaré subgroups that we obtain by this method would represent a symmetry group relevant to this intermediate state. Since the intermediate state is effectively stationary, there would be an asymptotic Poincaré symmetry subgroup of BMS that applies at least approximately

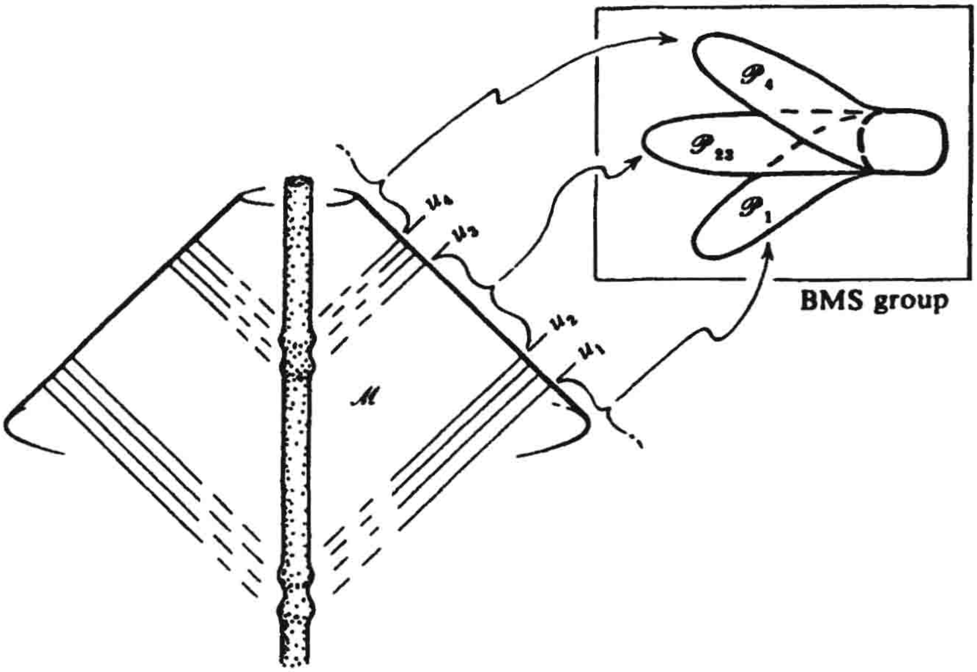


FIGURE 1. An isolated gravitating system emits two bursts of gravitational radiation, the system being effectively quiescent between the bursts. Three distinct Poincaré subgroups $\mathcal{P}_1$, $\mathcal{P}_{23}$, and $\mathcal{P}_4$ of the BMS group seem to be naturally singled out, having only their translation subgroup in common (Penrose and Rindler 1986).

and the angular momentum concept of relevance would be that which refers to the rotational part of this subgroup. What one appears to need is some concept of a good cut that applies at *all* retarded times, not just in the remote retarded past or future.

Very early in our studies we contemplated the condition $\sigma^0 = 0$ as a possible characterization of a good cut throughout the *whole* of $\mathscr{I}^+$, as this can indeed be used as the appropriate criterion in flat space-time **M**, or even in a general stationary space-time. But, this is much too strong a condition when radiation is present; and in general, no such cuts will exist for a curved nonstationary $\mathcal{M}$. The reason that this characterization is too strong can be understood in terms of the amount of functional freedom that is available in choosing a cut. This freedom is *one* real function of two real variables (since the cut is defined by a single real function on S^2, the u-value of the cut), whereas σ^0 is *complex*, and so its vanishing is *two* real conditions. A possible compromise procedure would be to demand that merely the *electric part* of σ^0 be set to zero over the cut. Indeed, this was tried by Ted and his collaborators (Aronson, Lind, Messmer, and Newman 1971). However, in general, no nontrivial subgroup of BMS arises as the group preserving this family of good cuts. In the example given of a system emitting two bursts of radiation, the "good cuts" in this sense, in the intermediate closely stationary state, would not (in general, at least) be translations of those either near i^0 or near i^+.

Ted's "crazy idea"—that led to the concept of $\mathcal{H}$-space—was to *complexify* $\mathcal{I}^+$ and to allow the cuts of this complexified null infinity, $\mathbb{C}\mathcal{I}^+$, to be complex. Since these complex cuts now have a complex u-value, there is the freedom of one *complex* number per point of S^2, i.e., of *two* real numbers per point. This is just the right amount of freedom to enable σ^0 to be killed off. Ted's new concept of *good cut* is now a *complex* cut of $\mathbb{C}\mathcal{I}^+$ for which $\sigma^0 = 0$.

A number of remarks should be made about this definition. First, there is a certain chiral asymmetry involved. It by no means follows from the choice $\sigma^0 = 0$ that we need also have $\tilde{\sigma}^0 = 0$, where $\tilde{\sigma}$ is the complexification of the *complex conjugate* $\bar{\sigma}$ of σ (so $\tilde{\sigma}$ is a complex quantity "freed" from its association with σ). It would have been an equally legitimate but different definition of a good cut to demand that $\tilde{\sigma}^0 = 0$, instead of $\sigma^0 = 0$. The two choices of definition are in fact space reflections of one another. Second, one must be somewhat careful about what is entailed in the use of the term "complexification." All the quantities that are involved in a complexification need to be real-*analytic*; but even so, it is not necessarily the case that the extensions into the complex can be carried very far. All that we can guarantee is that a certain "thickening" into a complex manifold, with a holomorphic (conformal) metric, exists; where the original $\mathcal{M} \cup \mathcal{I}^+$, with its real (conformal) metric, occurs as a "real slice." In the present situation, analyticity alone may not be sufficient to guarantee that complex good cuts exist. It is also necessary that the complex thickening can be chosen to be sufficiently extensive that the required good cuts can be located within it. This will always be the case if the original space-time is, in an appropriate sense, sufficiently close to the flat case (where "sufficiently close" would have to be taken in a sense that takes all the higher derivatives into account), but as far as I am aware, the conditions have not yet been precisely formulated.

If we do assume that the appropriate existence conditions are satisfied, then some remarkable conclusions follow, and we obtain Ted's extraordinary $\mathcal{H}$-space $\mathcal{H}$ as the *space* of good cuts for $\mathcal{M}$ defined in this way. Ted was able to show that $\mathcal{H}$ is *four dimensional* and that it possesses a complex Riemannian (holomorphic) metric defined by

$$g_{ab}V^aV^b = \left\{\oint(\delta V)^{-2}\,d\mathcal{S}\right\}^{-1}.$$

Here, δV represents the separation (in terms of complexified Bondi time) between the given cut and a neighboring one, δV being a function on S^2, the displacement in $\mathcal{H}$ from the corresponding point of $\mathcal{H}$ to its neighboring one being given by the (abstract-indexed) vector V^a; and $d\mathcal{S}$ is the surface element of S^2 (i.e., of the cut,) scaled in accordance with the choice of Bondi time. The $\oint$ is taken over the whole of the real S^2, or else over some 2-contour that is a deformation of this. It is easy to see that this definition satisfies all the invariance properties that would be required, but it is by no means obvious that the expression on the right defined in this way, is in fact *quadratic* in V^a, i.e., that we do get a complex Riemannian metric, as required, rather than some kind of Finsler metric.

2. Connections with Twistor Theory

My interest in Ted's $\mathcal{H}$-space was especially aroused because of some connections with twistor theory. (See Penrose and Rindler 1986, and Huggett and Tod 1985 for accounts of twistor theory; see Hansen, Newman, Penrose, and Tod 1978 for further details concerning the relation between $\mathcal{H}$-space and twistor theory.) We can define the appropriate projective twistor space $\mathbb{P}\mathcal{T}$ of $\mathcal{M}$—or its "dual" $\mathbb{P}\mathcal{T}^*$—in terms of spaces of *complex null geodesics* on complexified future null infinity $\mathbb{C}\mathcal{I}^+$. It should first be remarked that the complex 3-space $\mathbb{C}\mathcal{I}^+$ is separately fibred by two families of complex 2-planes called *α-planes* and *β-planes* (see Figure 2). The α-planes intersect the β-planes in the complex *generators* of $\mathbb{C}\mathcal{I}^+$. The null geodesics in $\mathbb{C}\mathcal{I}^+$ are of three kinds: those lying in β-planes but not in α-planes, called *α-curves*; those in α-planes but not in β-planes, called *β-curves*; and the generators of $\mathbb{C}\mathcal{I}^+$ themselves, lying in both. The *projective asymptotic twistor space* $\mathbb{P}\mathcal{T}$ is the space whose points represent α-curves, while the "dual" space $\mathbb{P}\mathcal{T}^*$ is the corresponding space of β-curves. It is a matter of sorting out one's conventions to decide which family consists of the α-curves and which the β-curves. (The α-curves have tangents of the form $\iota^A\pi^{A'}$, and the β-curves have tangents of the form $\lambda^A\iota^{A'}$, where $\iota^A\iota^{A'}$ are tangents to the generators.) The α-planes are distinguished from the β-planes by whether one chooses the stereographic coordinate ζ or the complexification (freed up version) $\tilde{\zeta}$ of its complex conjugate $\bar{\zeta}$ (in some order).

We shall find that it is the β-curves that are very closely associated with $\mathcal{H}$-space, while the corresponding use of the α-curves would lead us to the *conjugate* $\mathcal{H}$-space, defined by $\tilde{\sigma}^0 = 0$, rather than by $\sigma^0 = 0$. (It always seemed to be the case, after Ted and I had not seen each other for a while, that when we would get together again our conventions would have "flipped" relative to each other! This was a major contributing factor in our confusion at one stage when Ted thought that our results led to the erroneous conclusion the $\mathcal{H}$-space is *flat*. As we shall see, $\mathcal{H}$-space is a self-dual vacuum space, whereas conjugate $\mathcal{H}$-space is an anti-self-dual vacuum space. A vacuum space that is both at once must indeed be flat!)

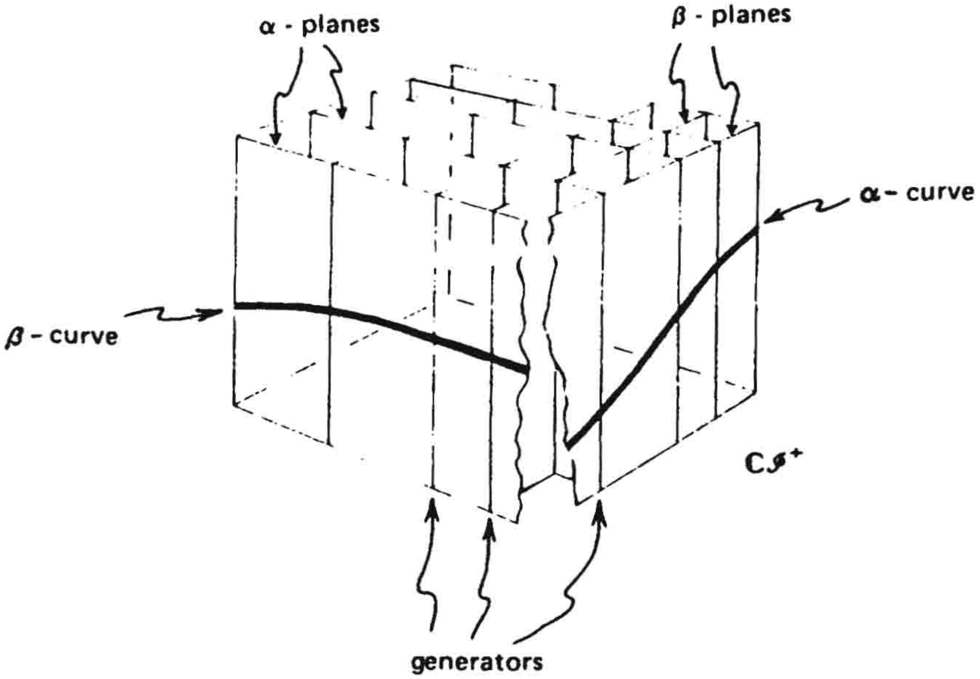


FIGURE 2. Complex null geodesics on $\mathbb{C}\mathcal{J}^+$. The β-curves lie in α-planes, and the α-curves lie in β-planes (Penrose and Ward 1980).

I had realized that the condition $\sigma^0 = 0$ could be interpreted as the fact that the good cuts of $\mathbb{C}\mathcal{J}^+$ are precisely the cuts that are ruled by β-curves (or else by α-curves, in accordance with the confusions referred to above), i.e., they are simply the surfaces that are swept out by a 1-parameter collection of β-curves. Thus, these good cuts have a clear "twistorial" interpretation as *curves* in $\mathbb{P}\mathcal{T}^*$. I had been very puzzled by the nature of these curves. What is it that restricts them to a 4-parameter family, so that the resulting $\mathcal{H}$-space, which represents the totality of such curves, will indeed be precisely four dimensional? I had imagined that there must be some differential equation needed in order to define the curves, and I tried to find out from Ted what this differential equation must be. I was startled to learn from him that there was *no* such differential equation; the family of curves is completely determined by the fact that each one exists *globally*—being a compact holomorphic curve (with the topology S^2) and belonging to the right topological family within $\mathbb{P}\mathcal{T}^*$!

The central insight from Ted's $\mathcal{H}$-space theory, which came as a complete surprise to me at the time (although I was subsequently able to relate it to what I already knew about the restricted nature of compact holomorphic curves in complex projective space), turned out to be the key ingredient for the emerging "curved" twistor theory. The so-called "nonlinear graviton construction" arose directly out of it. In this construction (Penrose 1976), we form a nonflat projective twistor space, which could be $\mathbb{P}\mathcal{T}^*$ or something a little more general, and the holomorphic curves in it belonging to the right topological family (in fact cross sections of the holomorphic fibration that $\mathbb{P}\mathcal{T}^*$ turns out to be) provide the "points" of a new space. This new space is always a *self-dual solution of the vacuum equations* (i.e., $R_{ab} = 0$, $C^*_{abcd} = iC_{abcd}$; equivalently, $\Phi_{ABA'B'}$, Λ, and Ψ_{ABCD} all vanish, according to the conventions of Penrose and Rindler 1984, 1986), and this construction yields the *general* self-dual solution (Penrose 1976). When this curved twistor space is indeed $\mathbb{P}\mathcal{T}^*$, we get $\mathcal{H}$-space, in accordance with the above. Thus, $\mathcal{H}$-space is always a self-dual vacuum![1] This holds whether or not $\mathcal{M}$ itself is vacuum near $\mathscr{I}^+$.

The definition of the metric for the "space-time" that arises from a general nonlinear graviton construction is not obviously equivalent to that which Ted had given for his $\mathcal{H}$-space, but the two can indeed be reduced to one another (see Hansen, Newman, Penrose, and Tod 1978 for a discussion of this and other matters concerning $\mathcal{H}$-space and its relation to twistor theory). The fact that $\mathcal{H}$-space is a self-dual space and not generally flat forbids its direct use in the construction of a "Poincaré group" that might be used to provide us with a concept of angular momentum. One can, however, give $\mathcal{H}$-space some sort of physical interpretation. For this, we consider the gravitational radiation field of our radiating space-time $\mathcal{M}$, and we decompose this field into its anti-self-dual and self-dual parts. This radiation field would be described by the asymptotic Weyl component Ψ^0_4 and its complex conjugate $\bar{\Psi}^0_4$. When we consider the complexified space-time $\mathbb{C}\mathcal{M}$, $\bar{\Psi}^0_4$ becomes $\tilde{\Psi}^0_4$, a quantity independent of Ψ^0_4. (In a quantized theory, where a positive frequency condition might be expected to hold, $\tilde{\Psi}_4^0$ would describe right-handed circularly polarized gravitons and Ψ_4^0 would describe left-handed ones.) A way of interpreting $\mathcal{H}$-space is that it is the sourceless complex vacuum space-time whose self-dual radiation field is precisely the same $\tilde{\Psi}_4^0$ as that for $\mathbb{C}\mathcal{M}$ and whose anti-self-dual radiation vanishes.

There is a way of obtaining a *flat* space-time using these ideas, namely, to take the *conjugate* $\mathcal{H}$-space of this $\mathcal{H}$-space, since this would kill off the self-dual gravitational radiation field as well as the anti-self-dual radiation field. One might consider that the complex Poincaré group of symmetries of this flat space might supply the required supertranslation-free "rotational symmetries" needed for our elusive definition of angular momentum. However, it turns out that all we get is the same Poincaré group that we had before, referring to the symmetries of the family of "asymptotically good cuts" near i^+. Thus, this procedure yields nothing essentially new in the understanding of angular momentum, and it suffers from the same problems as before.

Ingenious as the $\mathcal{H}$-space construction might be, it still seemed to shed no light on the definition of angular momentum.[2] One might seriously have wondered what $\mathcal{H}$-space and its twistorial connections could be good for. I recall having many long and tantalizing conversations with Ted in the early days of $\mathcal{H}$-space theory. We knew that we were profoundly intrigued by the idea, but we were not sure why; and we believed that we would never get anyone else to see the point of it or to be in any way interested in these concepts (not really knowing what the point was ourselves). Yet, we pressed on, Ted first and I trailing along afterward, in the belief that there was indeed something of importance to be uncovered. It is now clear that Ted was right. Let us catch a glimpse of what has indeed been revealed.

3. Twistors and Integrable Systems

The non-linear graviton construction, which, as we have seen, is the direct descendant of the $\mathcal{H}$-space conception, has the fundamental property that the twistor space ($\mathbb{P}\mathcal{T}$ or) $\mathbb{P}\mathcal{T}^*$ can be constructed entirely by the use of free holomorphic functions (see Penrose 1976, Section 6). Thus, there is a sense in which the general (anti-)self-dual Einstein equations form an *integrable system*. In fact, it is not at all direct to obtain, from the twistor space $\mathbb{P}\mathcal{T}^*$, the corresponding (anti-)self-dual "space-time." In order to locate the "space-time points," one needs to find the global holomorphic cross sections of $\mathbb{P}\mathcal{T}^*$ (as a holomorphic fibration). Indeed, this is what Ted's $\mathcal{H}$-space idea has told us we *must* do. The problem of finding these cross sections is a very nontrivial one, in general. Nevertheless, the term "integrable" seems not altogether unreasonable in these circumstances, and it has become commonly used in connection with the subsequent, and closely related, *Ward construction* for the solution of the (anti-)self-dual *Yang–Mills* equations.

The Ward construction (Ward 1977), which is, in a clear sense, $\mathcal{H}$-space's "grandchild," is illustrated in Figure 3. Now, we must consider *bundles* over regions of projective twistor space, and the cross sections of these bundles over the relevant holomorphic curves (over the lines of **PT** in the case of flat twistor space) provide the required (anti-)self-dual Yang–Mills connections. Again, general *free* holomorphic functions are all that are needed for the construction of the required Ward bundle, but if it is desired to obtain the Yang–Mills connections explicitly, it is necessary to find the required family of cross sections, just as in the case of the nonlinear graviton. This construction was used to great effect in the classic papers of Atiyah and Ward (1977) and Atiyah, Hitchin, Drinfeld, and Manin (1978), which effectively solved the instanton problem on S^4, and it has been generalized in various ways, e.g., by Buchdahl (1986) to solve the corresponding instanton problem on $\mathbf{CP}^2$. When symmetries are imposed on the Yang–Mills fields (or, equivalently, on the Ward bundle), we can obtain a "dimensional reduction" to obtain integrable systems of lower dimension (3 or 2). Ward (1990), Mason and Sparling (1989), and others have shown how to obtain most of the known integrable systems (Toda lattice, Sine–Gordon, Korteweg–de Vries, nonlinear Schrödinger) in this way. (These are $\mathcal{H}$-space's "great grandchildren"!)

An example of particular note for relativists is the family of stationary axisymmetric solutions of the Einstein vacuum equations (or Einstein–Maxwell or Einstein–Yang–Mills equations). No (anti-)self-dual conditions are imposed. Woodhouse and Mason (1988) were able to adapt Ward's (1983) procedures for this general situation in a novel and ingenious way (see also Fletcher and Woodhouse 1990). Here, one has two commuting continuous symmetries, and therefore two-dimensional orbits. The flat projective twistor space **PT** relevant to the problem also inherits these symmetries. Since **PT** has three complex dimensions, the space that is obtained by factoring out **PT** by the (still two dimensional) orbits is just one dimensional, i.e., this factor space is a *Riemann surface* $\mathcal{S}$. However, $\mathcal{S}$ is no ordinary Riemann surface, being *non-Hausdorff* in an essential way. The entire information of the axisymmetric space-time (and its fields thereon) is contained in the global nature of a holomorphic vector bundle over $\mathcal{S}$. In order for it to be possible to encode the sort of continuous information that is required in a bundle over a Riemann sur-

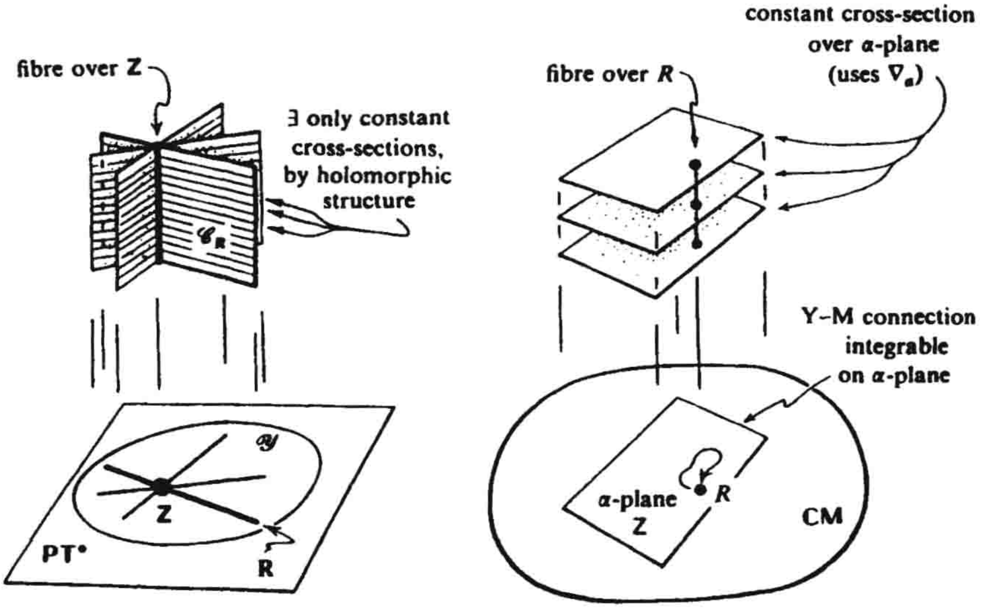

fibre over Z
∃ only constant cross-sections, by holomorphic structure
$\mathscr{C}_R$
$\mathscr{Y}$
Z
PT•
R
fibre over R
constant cross-section over α-plane (uses ∇_a)
Y–M connection integrable on α-plane
α-plane
Z
R
CM

FIGURE 3. The Ward construction for (anti-)self-dual Yang–Mills fields (Penrose and Rindler 1986).

face, it is necessary that this Riemann surface indeed be non-Hausdorff.

Non-Hausdorff complex manifolds have also had important roles to play in other areas of twistor theory. Most significantly, Bailey (1985, following an initial observation by Sparling concerning the Coulomb field's twistor space, cf. Penrose and Sparling 1990) has found that the description of *sources* in, for example, Maxwell theory, can be carried out in terms of (bundles over) non-Hausdorff twistor spaces. It seems likely that non-Hausdorffness in twistor theory will continue to play a very significant role. There is a clear connection between certain chaotic systems and non-Hausdorff Riemann surfaces (cf. Penrose 1990a for a way of describing Julia sets and the Mandelbrot set in these terms). In all these situations, it appears that one must adopt a particular philosophy with regard to the non-Hausdorffness that arises. Such complex manifolds are to be regarded as equivalent to one another if they can be obtained from each other by a succession of elementary "moves" involving the splitting apart or regluing of the non-Hausdorff leaves, where each of these moves does not, in some required sense, change the essential structure of the manifold as a whole. The precise formulation of this philosophy remains somewhat elusive, however.

4. Twistors and the General Einstein Equations

One of the major obstacles to progress in twistor theory continues to be the lack of a proper formulation of the full vacuum Einstein equations, i.e., without the (anti-)self-dual condition, in twistor terms. (See Penrose 1990b for an assessment of the significance and the status of this problem.) For a long time, this fundamental obstruction to progress has been addressed either in terms of the so-called "googly problem," or else in terms of "ambitwistors" instead of twistors. In the googly approach, one tries to find a way of "nonlinearizing" the procedure whereby a self-dual linearized Weyl curvature $\tilde{\psi}_{A'B'C'D'}$ can be obtained by contour integration from a twistor function $f(Z^\alpha)$ homogeneous of degree -6, this to be analogous to the way in which the above nonlinear graviton construction "nonlinearizes" the procedure whereby an anti-self-dual linearized Weyl curvature ψ_{ABCD} is obtained from a twistor function $g(Z^\alpha)$ homogeneous of degree $+2$ (see Penrose 1990b). There is actually a procedure for doing this (Penrose 1986), but it is not at all clear, as yet, whether this procedure is really what is required. Even if it can be made to work in a natural looking way, this procedure sheds little light on how one might *combine* the data for the self-dual and anti-self-dual parts of the Weyl curvature (which would amount to combining $\mathcal{H}$-space with the conjugate $\mathcal{H}$-space) to obtain a general vacuum space-time $\mathcal{M}$. More insight is needed.

With regard to ambitwistor approaches, where the space $\mathbf{P}\mathcal{A}$ of complex null geodesics—*ambitwistor space*—replaces projective twistor space as the primary complex manifold under consideration, there are other kinds of problems. (This is the space that in the flat case is represented by pairs of projective twistors and dual twistors Z^α, W_α, subject to to $Z^\alpha W_\alpha = 0$.) One of these problems is rooted in the fact that $\mathbf{P}\mathcal{A}$, being five dimensional, has more dimensions than the (complex) space-time $\mathbf{C}\mathcal{M}$ that it represents, so its deformations must be severely constrained, rather than being "free," as one might hope should be the case for a genuinely twistorial construction. Also, the direct ambitwistor approaches fit in less well with ideas of quantization than do the twistor approaches. (Twistor functions and nonlinear-graviton-type constructions are to be thought of as playing roles similar to those of wave functions, whereas ambitwistor deformations are more like classical phase-space descriptions; cf. Penrose 1990b.) Nevertheless, the ambitwistor constructions (LeBrun 1990; Baston and Mason 1987) ought to have a significant role to play in relation to whatever twistor procedures do finally prove to be most successful.

Recently, some new ideas (Mason and Penrose 1989) have arisen out of an attempt to relate twistor theory to Ashtekar's (1986) variables. Both twistor theory (as opposed to ambitwistor theory) and Ashtekar variables share the curious and perhaps significant property of being *chiral*. In twistor theory, we must make a choice as to whether the description in terms of *twistors* Z^α or *dual twistors* W_α is to be regarded as fundamental; with Ashtekar variables, one must correspondingly choose between the *primed* or *unprimed* spin connections, restricted to an initial space-like hypersurface Σ.

In order to make a correspondence between the two theories, we must first find the appropriate twistor concept that arises in relation to a hypersurface Σ (normally taken to be space-like). This is the concept of *hypersurface twistor*. The (dual) projective hypersurface twistor space $\mathbb{P}\mathcal{T}^*(\Sigma)$, for an analytic Σ in an analytic space-time $\mathcal{M}$, is the space of all the β-curves in the complexification $\mathbb{C}\Sigma$ of Σ, where a β-curve is defined as a holomorphic curve in $\mathbb{C}\Sigma$ whose tangent vectors have the form

$$\ell^a = \lambda^A \tilde{\lambda}^{A'},$$

where $\tilde{\lambda}^{A'} = N^{AA'}\lambda_{A'}$, N^a being normal to Σ (whence ℓ^a is tangential to λ^A), and is defined by the differential equation

$$\ell^a \nabla_a \lambda_B = 0.$$

(If $\mathcal{M}$ is conformally flat, or complex and merely conformally self-dual, so that it contains its full complement of β-planes, then the β-curves are just the intersections of its β-planes with $\mathbb{C}\Sigma$. Generally, the β-curves in $\mathbb{C}\Sigma$ may be thought of as what *would* be the intersections of $\mathbb{C}\Sigma$ with β-planes if $\mathbb{C}\Sigma$ had any β-planes!) The nonprojective hypersurface twistor space $\mathcal{T}^*(\Sigma)$ is obtained if we retain the field of λ_A spinors defined along the β-curves in addition to the β-curves themselves. (The definition of β-curves for $\mathbb{C}\mathcal{I}^+$ given previously turns out to be a special case of this.)

It is possible to construct a complex space-time $\mathcal{H}(\Sigma)$, called *Heaven on Earth*, by applying the nonlinear graviton construction to the (dual) hypersurface twistor space $\mathcal{T}^*(\Sigma)$. In general, this is conformally self-dual but not conformal to vacuum. Hence, its Ψ_{ABCD} vanishes, but its $\tilde{\Psi}_{A'B'C'D'}$, $\Phi_{ABA'B'}$, and Λ generally do not. We may think of $\mathcal{H}(\Sigma)$ as the evolution of initial data on Σ to produce not a vacuum space-time but a self-dual nonvacuum instead. This evolution (as opposed to the normal vacuum one) is conformally invariant. We may think of it as a construction producing the "space-time" that has *actual* β-planes meeting $\mathbb{C}\Sigma$ in its β-curves.

Correspondingly, there is also a *conjugate* Heaven on Earth, obtained by using α-curves in $\mathbb{C}\Sigma$ rather than β-curves (where a primed spinor $\pi_{A'}$ replaces λ_A in the definition). Thus, this conjugate space is the conformally *anti*-self-dual space obtained by applying the nonlinear graviton construction to the projective hypersurface twistor space $\mathbb{P}\mathcal{T}(\Sigma)$ of α-curves.

Let us see how a relation to Ashtekar variables can be made. The two sets of Ashtekar variables are (*1*) the "soldering form" on Σ, which translates between spinors λ_A (or rather their squares $\lambda_A\lambda_B$, which are equivalent to $\lambda^A\tilde{\lambda}^{A'}$) and tangent vectors ℓ^a to Σ, and (2) the restriction to tangent directions to Σ of the unprimed spinor connection of $\mathcal{M}$. The totality of this information is largely contained in the β-curve equation $\ell^a\nabla_a\lambda_B = 0$, i.e.,

$$\lambda^A N^{CA'}\lambda_{AA'}\lambda_B = 0,$$

but not quite. This equation falls short for two reasons. First, only the conformally invariant part of the connection is encoded, because under conformal rescaling, $\nabla_{AA'}\lambda_B$ would be replaced by

$$\nabla_{AA'}\lambda_B - \Upsilon_{BA'}\lambda_A,$$

where $\Upsilon_{AA'} = \Omega^{-1}\nabla_{AA'}\Omega$, with Ω being the conformal factor (see Penrose and Rindler 1984), and the β-curve equation is left unchanged by this. The β-curve equation determines only the symmetric part $\nabla_{A'(A}\lambda_{B)}$, whereas for Ashtekar variables, we need the whole of $\nabla_{A'A}\lambda_B$. Our first problem, therefore, is one of picking out a conformal factor. Second, we have specified no way in which the actual hypersurface $\mathbb{C}\Sigma$ is to be located within $\mathcal{H}(\Sigma)$. [$\mathbb{C}\Sigma$ is, indeed, canonically identified as a hypersurface within $\mathcal{H}(\Sigma)$, as well within $\mathbb{C}\mathcal{M}$.] $\mathcal{H}(\Sigma)$ is equivalent to $\mathbb{P}\mathcal{T}^*(\Sigma)$. Thus, our second problem can be reinterpreted as the following: how do we actually identify the (conformal) initial data at $\mathbb{C}\Sigma$ in terms of $\mathbb{P}\mathcal{T}^*(\Sigma)$?

Both of these shortcomings can be resolved in terms of a *bundle* $\mathcal{B}$ over $\mathbb{P}\mathcal{T}^*(\Sigma)$. The fibres of $\mathcal{B}$ are solutions of the equation

$$\lambda^A\tilde{\lambda}^{A'}\nabla_{AA'}\nu_B = \tilde{\lambda}^{A'}N_{BA'},$$

along each β-curve (and in fact, $\tilde{\lambda}^{A'}N_{BA'} = \frac{1}{2}N_aN^a\lambda_B$). There are two independent solutions of this equation (including a one-dimensional family with ν_B proportional to λ_B), so we have a rank-two bundle. It is not a vector bundle but an *affine* bundle since there is no natural zero for ν_B. We adopt the scaling rule

$$\nu_A \rightarrow \kappa^{-1}\nu_A \quad \text{under} \quad \lambda_A \rightarrow \kappa\lambda_A.$$

Then $\mathcal{B}$ can be legitimately regarded as a bundle over the projective space $\mathbb{P}\mathcal{T}^*(\Sigma)$ rather than over $\mathcal{T}^*(\Sigma)$. The total space $\mathcal{B}$ has then *five* (rather than six) dimensions.

The equation governing the behavior of ν_B along β-curves provides us with the needed remaining (nonconformal) part of the spin connection. The remaining piece of information that we need, in terms of $\mathbb{P}\mathcal{T}^*(\Sigma)$, is the hypersurface $\mathbb{C}\Sigma$. This can be identified as the family of *holomorphic cross sections* of $\mathcal{B}$ ($\mathcal{H}$-space's descendents again!) for which ν_A is proportional to λ_A, i.e., for which

$$\nu^A\lambda_A = 0.$$

For this to be part of the information contained in $\mathcal{B}$, we need to specify a little more structure, namely, we need the subspace for which ν_A is proportional to λ_A to be canonically identified with $\mathcal{B}$. One way of achieving this is in terms of a map:

$$\mathbf{X}: \mathcal{B} \to \mathbb{P}\mathcal{T}^*(\Sigma) \times \mathcal{C},$$

given by

$$\nu_A \to \nu^A\lambda_A$$

(noting that $\nu^A\lambda_A$ is constant along β-curves), and we look for cross sections over the inverse image of zero.

We can interpret this over the whole of $\mathcal{H}(\Sigma)$ (i.e., away from $\mathbb{C}\Sigma$) in terms of a (zero-weight) scalar quantity χ that, together with ν_A, can be defined on $\mathcal{H}(\Sigma)$ (using ∇ to denote derivatives within $\mathcal{H}(\Sigma)$, conformally scaled to have its scalar curvature zero) by the pair of equations

$$\lambda^A\nabla_{AA'}\nu_B = \nabla_{BA'}\chi$$

and

$$\Box\chi = 0,$$

where

$$\chi = 0 \quad \text{and} \quad \nabla_a\chi = N_a \quad \text{on} \quad \mathbb{C}\Sigma.$$

So far, this does not relate to Einstein's equations in any clear way. One way of achieving such a relationship is to introduce a further affine bundle $\mathcal{A}(\Sigma)$ over $\mathbb{P}\mathcal{T}^*(\Sigma)$ [or over $\mathcal{T}^*(\Sigma)$] that has rank four rather than two, this being a kind of "affinization" of ambitwistor space, and that also has a structure defined by a map:

$$\mathbf{Y}: \mathcal{A}(\Sigma) \to \mathbb{P}\mathcal{T}^*(\Sigma) \times \mathbb{C},$$

where now

$$Z^\alpha \to Z^\alpha W_\alpha + \chi,$$

the "fibre coordinate" now being Z^α (instead of ν_A), and where along the β-curve we can set up local twistor coordinates on $\mathbb{P}\mathcal{T}^*(\Sigma)$ as W_α. Here, Z^α is propagated along β-curves by local twistor transport, modified by the addition of the term $\nabla_{BA'}\chi$. We also have the scaling

$$Z^\alpha \to \kappa^{-1} Z^\alpha \quad \text{under} \quad W_\alpha \to \kappa W_\alpha.$$

A full coding of the constraint equations of the Einstein vacuum equations can be given in terms of $\mathcal{A}(\Sigma)$. (For most of the details, see Mason and Penrose 1989; also Penrose 1990b.)

Of course, even this does not provide us with a *global* description of the Einstein equations in twistor terms since we are tied to the hypersurface Σ (or $\mathbb{C}\Sigma$). What would seem to be required would be a twistor space $\mathcal{T}$ that applies to the whole space-time $\mathcal{M}$ (or $\mathbb{C}\mathcal{M}$) all at once. It does not seem possible to achieve this with a *projective* twistor space; we would need (at least) our putative four-complex-dimensional manifold $\mathcal{T}$ in order to encode all that is required. We might anticipate that the freedom of choosing different three-complex-dimensional submanifolds of $\mathcal{T}$ (necessarily nonglobal but, in an appropriate sense, cohomological) would correspond to the freedom of choosing the space-time hypersurface Σ (or $\mathbb{C}\Sigma$) within $\mathcal{M}$ (or $\mathbb{C}\mathcal{M}$). For this, our $\mathcal{T}$ would seem to have to be a deformed "blown-up" version of the flat twistor space **T**. One would hope that all the various (projective) hypersurface twistor spaces $\mathbb{P}\mathcal{T}(\Sigma)$ could then, in an appropriate sense, be located within $\mathcal{T}$.

Whatever definition $\mathcal{T}$ might turn out to have, it ought to have a well-defined interpretation in terms of the original (vacuum) space-time $\mathcal{M}$. Such an interpretation would have to be a highly nonlocal one. Only very recently a new suggestion toward achieving this has emerged (Penrose 1990c); it involves adapting a concept due to Andrew Hodges arising out of the theory of twistor diagrams: *elemental states* (Hodges 1990a; cf. Hodges 1990b for a survey of twistor diagram theory). These are entities that have a sheaf-cohomological (or worse?) space-time interpretation, and for the needed twistor space $\mathcal{T}$, the entities would be described in terms of helicity $\frac{3}{2}$ potentials. The Einstein vacuum equations are indeed the appropriate consistency condition for such potentials (Buchdahl 1962). It seems that this new twistor concept should also play a role in terms of the "charges" for helicity $\frac{3}{2}$ massless potentials, although this viewpoint presents some odd puzzles that are yet to be resolved. This general approach offers hope for a completely new and perhaps very promising restart to curved space twistor theory, which could have significant implications in a number of areas.

Whatever the status of this particular idea turns out to be, it is clear that the role of linear sheaf cohomology, and its nonlinear generalizations (such as occur in the nonlinear graviton and Ward constructions), has had a fundamental influence on twistor theory and other related areas. It was Ted's $\mathcal{H}$-space, at least for me, that started all this. My indebtedness to Ted, just for this one idea alone, is more than I can adequately express. Let us salute $\mathcal{H}$-space!

NOTES

[1] The $\mathcal{H}$ in Ted's $\mathcal{H}$-space stands for "heaven": "where the good Cohens (cones) go." When I first found an argument that I believed showed that $\mathcal{H}$-space is indeed vacuum, I sent Ted a telegram (from Trieste) saying simply "HEAVEN IS EMPTY." I remember this message being very puzzling to the people in the office from which I sent the telegram!

[2] More recently, using a quite different but still twistor-based set of ideas, it seems to have been possible to make some genuine headway with the angular momentum problem (Penrose 1982; cf. also Penrose and Rindler 1986, Section 9.9). Instead of looking for good cuts, one *compensates* for the lack of "goodness" on any particular cut.

It should be mentioned that there is the alternative approach to the problem of angular momentum, promoted particularly by Ashtekar and coworkers (cf. Ashtekar and Streubel 1981), which amounts to giving up the search (at least temporarily) for a specific 6-component concept of angular momentum from among the infinite-parameter set of "super-angular momenta." However, it seems to me that this is rather too defeatist, especially in the context of examples such as that illustrated in Figure 1, where there should be a way of singling out the correct components, at least in the intermediate state.

REFERENCES

Aronson, B., Lind, R., Messmer, J., and Newman, E.T. (1971). "A note on Asymptotically Flat Spaces." *Journal of Mathematical Physics* 12: 2462–2467.

Ashtekar, A. (1986). "New Variables for Classical and Quantum Gravity." *Physical Review Letters* 57: 2284.

Ashtekar, A., and Streubel, M. (1981). "Symplectic Geometry of Radiative Modes and Conserved Quantities at Null Infinity." *Proceedings of the Royal Society* (London) A376: 585–607.

Atiyah, M.F., Hitchin, N.J., Drinfeld, V.G., and Manin, Yu.I. (1978). "Construction of Instantons." *Physics Letters* A65: 185–187.

Atiyah, M.F., and Ward, R.S. (1977). "Instantons and Algebraic Geometry." *Communications in Mathematical Physics* 55: 111–124.

Bailey, T.N. (1985). "Twistors and Fields with Sources on Worldlines." *Proceedings of the Royal Society* (London) A397: 143–155.

Bailey, T.N., and Baston, R.J., eds. (1990). "Twistors in Mathematics and Physics." London and New York: Cambridge University Press.

Baston, R.J., and Mason, L.J. (1987). "Conformal Geometry, the Einstein Equations and Spaces of Complex Null Geodesics." *Classical and Quantum Gravity* 4: 815–826

Bondi, H. (1960). "Gravitational Waves in General Relativity." *Nature* (London) 186: 535.

Bondi, H., van der Burg, M.G.J., and Metzner, A.W.K. (1962). "Gravitational Waves in General Relativity. VII. Waves from Axi-Symmetric Isolated Systems." *Proceedings of the Royal Society* (London) A269: 21–52.

Buchdahl, H.A. (1962). "On the Compatibility of Relativistic Wave Equations in Riemann Spaces." *Nuovo Cimento* 25: 486–496.

Buchdahl, N.P. (1986). "Instantons on $\mathbb{CP}^2$." *Journal of Differential Geometry* 24: 19–52.

Fletcher, J., and Woodhouse, N.M.J.W. (1990). "Twistor Characterization of Stationary Axisymmetric Solutions of Einstein's Equations." In *Twistors in Mathematics and Physics*, Bailey and Baston, eds. London and New York: Cambridge University Press, pp. 260–282

Hansen, R.O., Newman, E.T., Penrose, R., and Tod, K.P. (1978). "The Metric and Curvature Properties of $\mathcal{H}$-Space." *Proceedings of the Royal Society* (London) A363: 445–468.

Hodges, A.P. (1990a). "Elemental States." In *Further Advances in Twistor Theory, Vol. 1: The Penrose Transform and Its Applications* (Pitman Notes in Mathematics 231), Mason and Hughston, eds. Essex, U.K.: Longman, Harlow; and New York: Wiley, pp. 308–310. [From *Twistor Newsletter* 22 (1986).]

—— (1990b). "Twistor Diagrams and Feynman Diagrams." In *Twistors in Mathematics and Physics*, Bailey and Baston, eds. London and New York: Cambridge University Press, pp. 339–366.

Huggett, S.A., and Tod, K.P. (1985). "An Introduction to Twistor Theory." London and New York: Cambridge University Press.

Ko, M., Ludvigsen, M., Newman, E.T., and Tod, K.P. (1981). "The Theory of $\mathcal{H}$-space." *Physics Reports* 71: 51–139.

LeBrun, C.R. (1990). "Twistors, Ambitwistors and Conformal Gravity." In *Twistors in Mathematics and Physics*, Bailey and Baston, eds. London and New York: Cambridge University Press, pp. 71–86

Mason, L.J., and Hughston, L.P., eds. (1990). *Further Advances in Twistor Theory, Vol. 1: The Penrose Transformation and Its Applications* (Pitman Notes in Mathematics 231), Essex, U.K.: Longman, Harlew; and New York: Wiley.

Mason, L.J., and Penrose, R. (1989). "A Twistorial Approach to the Full Vacuum Equations." *Twistor Newsletter* 29: 1–5.

Mason, L.J., and Sparling, G.A.J. (1989). "Non-Linear Schrödinger and Korteweg–de Vries are Reductions of Self-Dual Yang–Mills." *Physical Review Letters* A137: 29–33.

Newman, E.T. (1976). "Heaven and Its Properties." *General Relativity and Gravitation* 7: 107–111.

Newman, E.T., and Penrose, R. (1966). "A Note on the Bondi–Metzner–Sachs Group." *Journal of Mathematical Physics* 7: 863–870.

Penrose, R. (1963). "Asymptotic Properties of Fields and Space-Times." *Physical Review Letters* 10: 66–68.

—— (1974). "Relativistic Symmetry Groups." In *Group Theory in Non-Linear Problems* A.O. Barut, ed. Dordrecht: Reidel.

—— (1976). "Nonlinear Gravitons and Curved Twistor Theory." *General Relativity and Gravitation* 7: 31–52.

—— (1982). "Quasi-Local Mass in General Relativity." *Proceedings of the Royal Society* (London) A381: 53–63.

—— (1986). "Exponentiating a Relative H^2." *Twistor Newsletter* 21: 2–4.

—— (1990a). "Non-Hausdorff Riemann Surfaces and Complex Dynamical Systems." In *Further Advances in Twistor Theory, Vol. 1: The Penrose Transform and Its Applications* (Pitman Notes in Mathematics 231), Mason and Hughston, eds. Essex, U.K.: Longman, Harlow; and New York: Wiley, pp. 375–379.

—— (1990b). "Twistor Theory after 25 Years—Its Physical Status and Prospects." In *Twistors in Mathematics and Physics*, Bailey and Baston, eds. London and New York: Cambridge University Press, pp. 1–29.

—— (1990c). "Twistor Theory for Vacuum Space-Times: A New Approach." *Twistor Newsletter* 31: 6–8.

Penrose, R., and Rindler, W. (1984). "Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields." London and New York: Cambridge University Press.

—— (1986). "Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry." London and New York: Cambridge University Press.

Penrose, R., and Sparling, G.A.J. (1990). "The Anti-Self-Dual Coulomb Field's Non-Hausdorff Twistor Space." In *Further Advances in Twistor Theory, Vol. 1: The Penrose Transform and Its Applications* (Pitman Notes in Mathematics 231), Mason and Hughston, eds. Essex, U.K.: Longman, Harlow; and New York: Wiley, pp. 348–349. [from *Twistor Newsletter* 9 (1979)].

Penrose and Ward (1980). "Twistors for Flat and Curved Space-Time." In *General Relativity and Gravitation: One Hundred Years after the Birth of Albert Einstein*, A. Held, ed. New York: Plenum, Ch. 7: pp. 283–328.

Sachs, R.K. (1962a). "Gravitational Waves in General Relativity. VIII. Waves in Asymptotically Flat Space-Time." *Proceedings of the Royal Society* (London) A270: 103–126.

——— (1962b). "Asymptotic Symmetries in Gravitational Theory." ***Physical Review*** 128: 2851–2864.

Ward, R.S. (1977). "On Self-Dual Gauge Fields." ***Physics Letters*** A61: 81–82.

——— (1983). "Stationary and Axi-Symmetric Spacetimes." ***General Relativity and Gravitation*** 15: 105–109.

——— (1990). "Integrable Systems in Twistor Theory." In ***Twistors in Mathematics and Physics***, Bailey and Baston, eds. London and New York: Cambridge University Press, pp. 246–259.

Woodhouse, N.M.J., and Mason, L.J. (1988). "The Geroch Group and Non-Hausdorff Twistor Spaces." ***Nonlinearity*** 1: 73–114.

209

COMMENTS ON 'PENROSE AND MATHEMATICAL ABILITY'

W. S. Robinson, with comments by Roger Penrose. *Analysis (Oxford)* **52** (1992), no. 2, 80–8

As with Chapter 206, we do not have Robinson's original criticisms here, but it is possible to infer much about them from my own responses. My 'new book' referred to here was subsequently published as *Shadows of the Mind* (Oxford University Press, 1994).

COMMENTS ON 'PENROSE AND MATHEMATICAL ABILITY'

By Roger Penrose

I SHOULD like to make a few comments concerning William Robinson's critique (in Analysis, this issue) of my use of Gödel's theorem as an argument for a non-computable ingredient in conscious thinking. The first point I should make is that Professor Robinson's conclusion [S_2], when it comes down to it, is basically one that I had already considered. He concludes that, consistently with the Gödel argument, there could be some complicated algorithm *A* that mathematicians *unconsciously* use to decide mathematical questions (his [S_2]), while I suggest mathematicians might use an algorithm of such complexity that its validity can never be known to us. As far as I can see, we differ, essentially, only in that he takes this to be a plausible ('intelligible') possibility whereas I argue that it does not appear to concur with the way that mathematicians actually do mathematics. I am happy to agree that the whole topic needs further discussion and, in a new book that I am currently writing, I propose to address all the relevant arguments that have come to my attention in a good deal of detail. I think that Professor Robinson makes a very useful clarification in distinguishing between a putative algorithm that a mathematician might *consciously* use (his [S_1]) from one that he or she might *unconsciously* use ([S_2]). To my way of thinking, since we are here concerned with the perceiving and acceptance of mathematical truth – and *not* with the various things that might guide a mathematician in fruitful directions of discovery and inspiration – it is the putative conscious algorithm [S_1] that is of interest here (and I do not see that it is 'very obviously false' for a *single* discipline like arithmetic).

In my book I used the term 'valid' for a system (say for arithmetic) that yields only mathematical truths and no falsehoods. I have since learned that the approved term that logicians use for this is 'sound'. A belief in the soundness of a formal system entails a belief in its consistency. I thus claim that it is unreasonable for a mathematician (consciously) to believe in the results deduced from a formal system without also believing in its consistency. Therefore (by Gödel's theorem) no formal system (and hence no algorithm) can encapsulate all the conscious beliefs of a reasonable mathematician. So I don't believe in [S_1] either.

Finally, I should comment that, of course, there is no point in distinguishing the use of a *single* algorithm from the use of any finite collection of algorithms (or the use of an algorithmically 'self-improving' algorithm). These things are all included in what is meant by 'algorithmically', which means what can be achieved on a *single* (idealized) algorithmic computer.

Mathematical Institute,
University of Oxford,
24–29 St. Giles', Oxford OX1 3LB

210

THE MOST GENERAL {2,2} SELF-DUAL VACUUM: A GOOGLY APPROACH

Twistor Newsletter **34** (1992). With L. Haslehurst. Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/ CRC, London, 2001

I do not believe that my D. Phil student Linda Haslehurst ever completed any write-up of her thesis, so this article may have to be taken as the documentation of her work.

194 The most general {2,2} self-dual vacuum: A googly approach. *Twistor Newsletter* **34** (1992). With L. Haslehurst. Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001

§III.4.23 **The most general $\{2,2\}$ self-dual vacuum: a googly approach** *by L. Haslehurst and R. Penrose* (𝕋ℕ 34, May 1992)

Twistor functions of homogeneity degree -6 can be used to generate curved dual twistor spaces and, thence, self-dual (complex) vacuum "space-times" $\mathscr{M}$. There are some plausible advantages in proceeding in this "googly" fashion (Penrose §III.4.11, §III.4.12 and §III.4.17). Any symmetry that the twistor function f possesses will be carried over to a symmetry of $\mathscr{M}$, and there is a natural specialization to the linearized limit. Moreover, since in linear theory a pole-multiplicity m in the -6 homogeneity function f is reflected in the (linearized) Weyl curvature possessing a $(5-m)$-fold principal null direction Penrose (1969), we may anticipate something similar occurring in the fully nonlinear self-dual case. We have been able to carry through this programme explicitly in the case when f is the inverse cube of a generic quadratic form, and find that, indeed, the resulting self-dual $\mathscr{M}$ has $\{2,2\}$ Weyl tensor, being the self-dual specialization of the Plebanski-Demianski solution (Plebanski & Demianski 1976), which is itself a NUT-type generalization of Levi-Civitá's C-metric. (Our solution belongs to the class described by Tod & Ward (1979), and also one by Fette, Janis & Newman (1976).

The general quadratic form $A_{\alpha\beta}Z^\alpha Z^\beta$ in the twistor variable Z^α can be put in the form

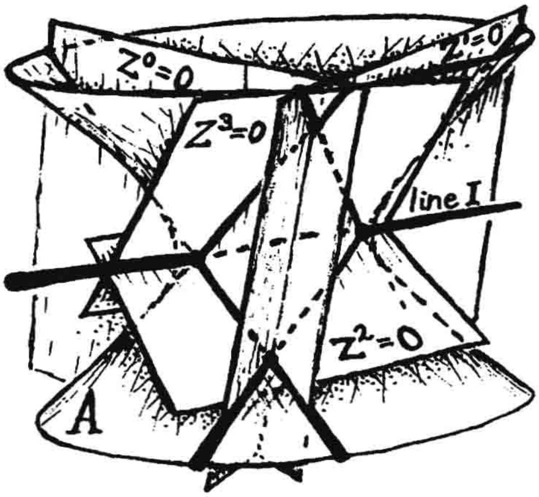
$$A_{\alpha\beta}Z^\alpha Z^\beta = Z^0Z^1 + a^2Z^2Z^3 = \omega^0\omega^1 + a^2\pi_{0'}\pi_{1'}$$

$$= \frac{1}{2}(Z^0 \ldots Z^3)\left(\begin{array}{cc:cc} & 1 & & \\ 1 & & & \\ \hdashline & & & a \\ & & a & \end{array}\right)\begin{pmatrix} Z^0 \\ \vdots \\ Z^3 \end{pmatrix}$$

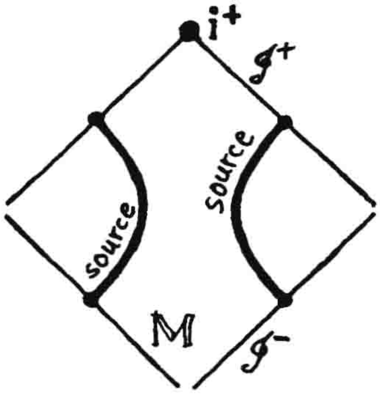
by a suitable choice of twistor coordinates ($Z^0 = \omega^0, Z^1 = \omega^1, Z^2 = \pi_{\omega'}, Z^3 = \pi_{1'}$) and our twistor function can be taken as

$$f = \left(\frac{k}{A_{\alpha\beta}Z^\alpha Z^\beta}\right)^3 = \frac{k^3}{(Z^0Z^1 + a^2Z^2Z^3)^3}.$$

This function has just 3-fold poles lying along the quadric $\boldsymbol{A}$ given by $A_{\alpha\beta}Z^\alpha Z^\beta = 0$. The line $\boldsymbol{I}$ is given by $Z^2 = Z^3 = 0$, and we have taken the generic case whereby $\boldsymbol{I}$ meets $\boldsymbol{A}$ in two distinct points.



If we put f into a twistor integral, we derive the linearized Weyl spinor of a linearized self-dual C-metric (uniformly accelerating NUT-like source). We have included "a" as an explicit parameter so as to allow for the limit $a \to 0$, which gives an elementary state in linear theory and the Eguchi-Hanson-Sparling-Tod solution in the full theory (§III.1.2, §III.1.3 and §III.4.2).



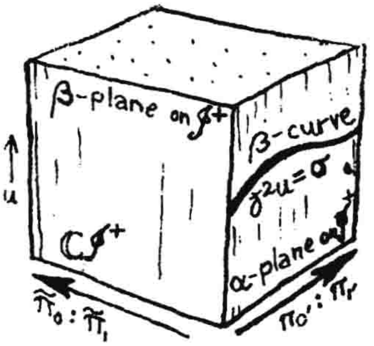
We do not recover the linearized or full Schwarzschild-NUT specialization with our particular coordinate choice, according to which the line I would lie *on* A and the twistor function could be written in the form $k(Z^0Z^2 + Z^1Z^3)^{-3}$. (This solution might encounter difficulties with a googly-type approach in any case since its $\mathbb{C}\mathscr{I}^+$ "looks" flat.) We do recover *flat* space, however, taking $k \to 0$.

The procedure is first to derive a "Bondi" σ $(= \sigma^0)$ on $\mathbb{C}\mathscr{I}^+$ (identified with the

196 The most general {2,2} self-dual vacuum: A googly approach. *Twistor Newsletter* **34** (1992). With L. Haslehurst. Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. 2001

standard $\mathbb{C}\mathscr{I}^+$ for $\mathbb{CM}$ by matching the two at i^+) according to the formula

$$\frac{\partial^2\sigma}{\partial\omega^A\partial\omega^B} = \oint \tilde{\pi}_A\tilde{\pi}_B f(\omega^A + \lambda\tilde{\pi}^A, \pi_{A'})d\lambda.$$



β-plane on $\mathscr{I}^+$
β-curve
$\eth^2 u = \sigma$ on $\mathscr{I}^+$
u
$\mathbb{C}\mathscr{I}^+$
α-plane on $\mathscr{I}^+$
$\tilde{\pi}_0 : \tilde{\pi}_1$
$\pi_{0'} : \pi_{1'}$

We then solve the Newman equation "$\eth^2 u = \sigma$" to derive the dual twistor lines (β-curves) on $\mathbb{C}\mathscr{I}^+$. The space of these lines provides us with a curved dual projective twistor space $\mathbb{P}\mathscr{T}^*$, and we find patching relations for $\mathscr{T}^*$ explicitly. Holomorphic cross-sections of $\mathscr{T}^*$, i.e., "good cuts" of $\mathbb{C}\mathscr{I}^+$ give us the points of $\mathcal{M}$, and these can also be found explicitly.

A standard u-coordinate for $\mathbb{C}\mathscr{I}^+$ for $\mathbb{M}$ can be given as

$$u = \mathrm{i}\tilde{\omega}^{A'}\pi_{A'} = -\mathrm{i}\,\tilde{\pi}_A\omega^A$$

(where the flat space twistor $(\omega^A, \pi_{A'})$ and dual twistor $(\tilde{\pi}_A, \tilde{\omega}^{A'})$ are incident and define a null ray in $\mathbb{CM}$ meeting $\mathbb{C}\mathscr{I}^+$ in a point with coordinate u). We can specialize either to $\pi_{0'} = 1$ or $\pi_{1'} = 1$ and either to $\tilde{\pi}_0 = 0$ or $\tilde{\pi}_1 = 1$, if desired, but we prefer to consider u as a homogeneity $(1,1)$ weighted expression and define a $(0,0)$ quantity ϱ by

$$\frac{u^2}{\pi_{0'}\pi_{1'}\tilde{\pi}_0\tilde{\pi}_1} = \frac{1}{\varrho}(\varrho + a^2)^2$$

Note that u is invariant under $\varrho \mapsto a^4/\varrho$.

It turns out that with our choice of f, the solution of Newman's equation for the β-curves on $\mathbb{C}\mathscr{I}^+$ is given by

$$F\frac{\pi_{1'}}{\pi_{0'}} = (w + P)\left(\frac{w + Q}{w - Q}\right)^{a^2/Q}$$

— or the same with $\pi_{0'}$ and $\pi_{1'}$ interchanged — where w is related to ϱ by

$$\varrho = \frac{w^2 - Q^2}{2(w+P)},$$

Q being a constant

$$Q^2 = a^4 - 8k^3,$$

and F and P being functions of $\tilde{\pi}_0$, $\tilde{\pi}_1$ (weights $(0,0)$ each).

We have ϱ unchanged under

$$w \mapsto \widehat{w} = -\frac{Pw + Q^2}{w+P}.$$

This leads to the patching relation $\widehat{W}_0 = W_0$, $\widehat{W}_1 = W_1$, and

$$\widehat{W}_2 = \left(1 + Q\frac{W_0W_1}{W_2W_3}\right)^{\frac{1}{2}\left(1-\frac{a^2}{Q}\right)} W_2, \quad \widehat{W}_3 = \left(1 + Q\frac{W_0W_1}{W_2W_3}\right)^{\frac{1}{2}\left(\frac{a^2}{Q}-1\right)} W_3$$

where

$$W_0 = \tilde{\pi}_0, \quad W_1 = \tilde{\pi}_1, \quad P - Q = \frac{2W_2W_3}{W_0W_1}, \quad \frac{W_2}{W_3} = \frac{Q-P}{F}, \quad \frac{\widehat{W}_2}{\widehat{W}_3} = \frac{\widehat{F}}{Q-P}$$

and

$$\widehat{P} = P, \qquad \widehat{F} = (P^2 - Q^2)\left(\frac{P+Q}{P+Q}\right)^{a^2/Q} \cdot \frac{1}{F}$$

(opposite choice of $\pi_{0'}$, $\pi_{1'}$ in β-curve equation for $\widehat{F}$). The patching relation between $\widehat{W}_{\underline{\alpha}}$ and $W_{\underline{\alpha}}$ defines our required dual twistor space $\mathscr{T}^*$. We have $d\widehat{W}_2 \wedge d\widehat{W}_3 = dW_2 \wedge dW_3$ on constant W_0, W_1, as required.

Note that $\widehat{W}_2\widehat{W}_3 = W_2W_3$, and that the two commuting symmetries

$$W_0 \mapsto \alpha W_0, \quad W_1 \mapsto \alpha^{-1}W_1 \quad \text{and} \quad W_2 \mapsto \beta W_2, \quad W_3 \mapsto \beta^{-1}W_3$$

hold for $\mathscr{T}^*$ and therefore give two commuting Killing vector symmetries for $\mathcal{M}$. The holomorphic cross-sections for $\mathscr{T}^*$ can be found explicitly. The Weyl curvature for $\mathcal{M}$ is indeed self-dual $\{2,2\}$, the solution being of Plebanski-Demianski type.

It is of interest to examine the weak field limit, where $Q = a^2 + \varepsilon$ (ε small). We find

$$\widehat{W}_2 = W_2\left(1 + \frac{\varepsilon}{2a^2}\log\left(1 + a^2\frac{W_0W_1}{W_2W_3}\right)\right)$$
$$\widehat{W}_3 = W_3\left(1 - \frac{\varepsilon}{2a^2}\log\left(1 + a^2\frac{W_0W_1}{W_2W_3}\right)\right)$$

with $\widehat{W}_0 = W_0$, $\widehat{W}_1 = W_1$. This comes from the standard formula

$$\widehat{W}_\alpha = W_\alpha + \varepsilon I_{\alpha\beta} \frac{\partial}{\partial W_\beta} \tilde{f}(W)$$

where

$$2\tilde{f}(W) = B \log B - C \log C - D \log D$$

with

$$B = W_0 W_1 + \frac{1}{a^2} W_2 W_3, \quad C = \frac{W_2 W_3}{a^2}, \quad D = W_0 W_1,$$

so $B = C + D$, ensuring $+2$ homogeneity. The term $B \log B$ is the kind of thing one expects from twistor transform considerations in relation to our original $f(z)$, since $B^{\alpha\beta}$ is the inverse of $A_{\alpha\beta}$, where $B = B^{\alpha\beta} W_\alpha W_\beta$.

References

Fette, C.W., Janis, A.I. & Newman, E.T. (1976) Algebraically special $\mathscr{H}$-spaces, J. Math. Phys., **17**, no. 5, 660–664.

Penrose, R. (1969) Solutions of the zero-rest-mass equations, J. Math. Phys., **10**, 38–39.

Plebanski, J.F. & Demianski, M. (1976) Rotating, charged, and uniformly accelerating mass in general relativity, Ann. of Phys., **98**, no. 1, 98–127.

Tod, K.P. & Ward R.S. (1979) Self-dual metrics with self-dual Killing vectors, Proc. Roy. Soc. Lond., **A 368**, no 1734, 411–427.

211

ON IMPULSIVE GRAVITATIONAL WAVES

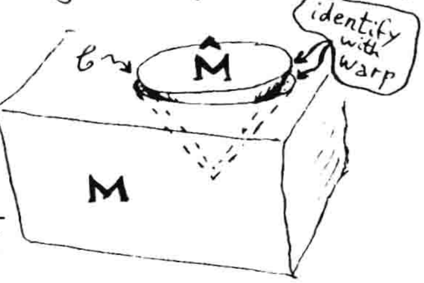
Twistor Newsletter **34** (1992). With Y. Nutku

Yavuz Nutku and I had intended to write a substantial account of this work, which includes a description of an encounter between two cosmic strings with the emission of an impulsive burst of gravitational radiation. However, we never found an occasion to complete the detailed write-up of this material. The closely related paper by Gleiser and Pullin (referred to at the end) had not been initially known to us.

On Impulsive Gravitational Waves

The general pure δ-function solution of the Einstein vacuum equations was described in Penrose (1972), using a "scissors and paste" construction. Such waves can have plane or spherical wave fronts, the former being limiting cases of the latter. We discuss only the spherical case here. The scissors and paste description can be given by matching a region **M** of Minkowski space outside a (say future) light cone $\mathscr{C}$, with metric

$$ds^2 = 2\,du\,dv - 2u^2 d\zeta d\bar{\zeta}$$



to another Minkowskian region $\hat{\mathbf{M}}$ inside $\mathscr{C}$, with metric

$$ds^2 = 2\,d\hat{u}\,d\hat{v} - 2\hat{u}^2 d\hat{\zeta} d\bar{\hat{\zeta}},$$

where the metric identification ("warp") at $\mathscr{C}$ is given by

$$\begin{aligned}\hat{v} &= 0 = v\\ \hat{\zeta} &= f(\zeta)\\ \hat{u} &= u/|f'(\zeta)|,\end{aligned} \qquad *$$

f being holomorphic, apart from at its singular regions, which may be thought of as singular "wires" on $\mathscr{C}$. The Einstein vacuum equations are satisfied across $\mathscr{C}$, along which there is a δ-function in the Weyl curvature.

It is of interest to note that the (u, ζ) transformation can be described neatly as a non-linear holomorphic transformation of the spin-space, at the vertex **O** of $\mathscr{C}$, preserving

$$\boxed{\xi_A\, d\xi^A} \qquad \text{i.e. both of } \lfloor d\xi_A \wedge d\xi^A \rfloor \text{ and } \left\lfloor \xi^A \frac{\partial}{\partial \xi^A} \right\rfloor.$$

Here, the position vector of a point on the cone is given by $\xi^A \bar{\xi}^{A'}$ $(= x^a)$, which in coordinates is $(\frac{1}{\sqrt{2}}\,x)$

$$\begin{bmatrix} u & \bar\zeta u \\ \zeta u & \zeta\bar\zeta u \end{bmatrix} = \begin{bmatrix} \xi^0\bar\xi^{0'} & \xi^0\bar\xi^{1'} \\ \xi^1\bar\xi^{0'} & \xi^1\bar\xi^{1'} \end{bmatrix},$$

so $\ast$ becomes $\xi^A \mapsto \hat\xi^A$ according to

$$\xi^0 \mapsto \hat\xi^0 = \xi^0\,(f'(\zeta))^{-\frac12}$$
$$\xi^1 \mapsto \hat\xi^1 = \xi^0 f(\zeta)\,(f'(\zeta))^{-\frac12}$$

where

$$\zeta = \xi^1/\xi^0$$

i.e.

$$\zeta \mapsto \hat\zeta = f(\zeta) \qquad\qquad \text{where } \begin{pmatrix}\xi^0\\ \xi^1\end{pmatrix} = \begin{pmatrix}\eta^{\frac12}\\ \zeta\eta^{\frac12}\end{pmatrix}.$$
$$\eta \mapsto \hat\eta = \eta/f'(\zeta)$$

This preserves

$$\xi_A\, d\xi^A = \xi^0 d\xi^1 - \xi^1 d\xi^0 = (\xi^0)^2\, d\left(\frac{\xi^1}{\xi^0}\right) = \eta\, d\zeta.$$

This fact is closely related to the Hamiltonian nature of the twistor transformation between **M**'s and $\hat{\mathbf{M}}$'s twistor spaces, that was noted in Penrose & MacCallum (1972).

The metric on the entire space $\mathcal{M} = \mathbf{M} \cup \mathcal{C} \cup \hat{\mathbf{M}}$ can be described as a $\mathbf{C}^0$ metric form

$$ds^2 = 2\,du\,dv - 2\left|u\,d\bar\zeta + v\{h;\zeta\}\,d\zeta\right|^2$$

where $\{\;;\;\}$ stands for the Schwarzian derivative

$$\{h;\zeta\} = -\frac12\left(\frac{h_{\zeta\zeta\zeta}}{h_\zeta} - \frac32\left(\frac{h_{\zeta\zeta}}{h_\zeta{}^2}\right)^2\right).$$

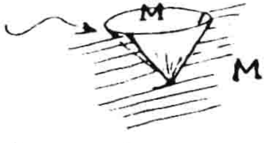
The curvature is defined by the only surviving Weyl component

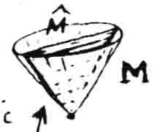
$$\Psi_4 = \frac1u\,\{h;\zeta\}_\zeta\,\delta(v)$$

in a suitable spin frame (o^A pointing along generators of $\mathcal{C}$).
(Nutku 1990)

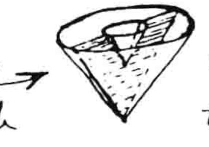
A simple example is given by a "snapping

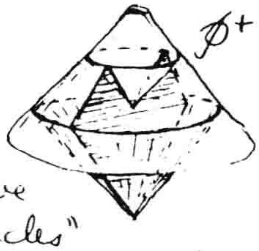
coon scribed, in a different C^0 way

by Gleiser & Pullin 1989), where $f(\zeta)=\zeta^{1+\varepsilon}$. Here, the "wires" at the north and south poles arise from a deficit-angle identification for a cosmic string in **M** which snaps, in $\mathcal{M}$, emitting a gravitational wave along $\mathscr{C}$ 

— or else in $\hat{\mathbf{M}}$, 

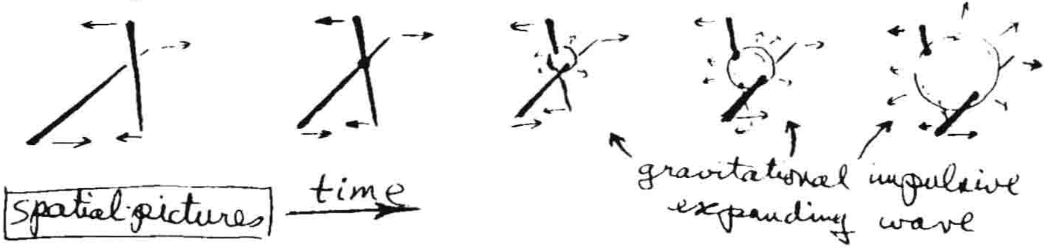
where here the cosmic string is created by the gravitational wave along $\mathscr{C}$ — or, analogously, we could snap or create a so-called "rotating" cosmic string if we allow ε to be complex.

An interesting consideration is that of energy balance, e.g. in the situation , where two string segments separate with the speed of light, having previously been joined as one string that was created with the emission of the first of the two gravitational wave bursts. Here the gravitational energy in the waves is infinite owing to an angular divergence, though the time-integral of energy flux along each generator of $\mathscr{I}^+$ is proportional to the length of the string segments. To make sense of all this, we must observe that the string segments have "particles" at their ends, where for a string of positive tension (and positive mass) the leading particle has a negative mass that decreases (becomes more negative) with time and the trailing one has a positive mass that increases with time; for a string of negative tension (i.e. positive pressure and hence negative mass)

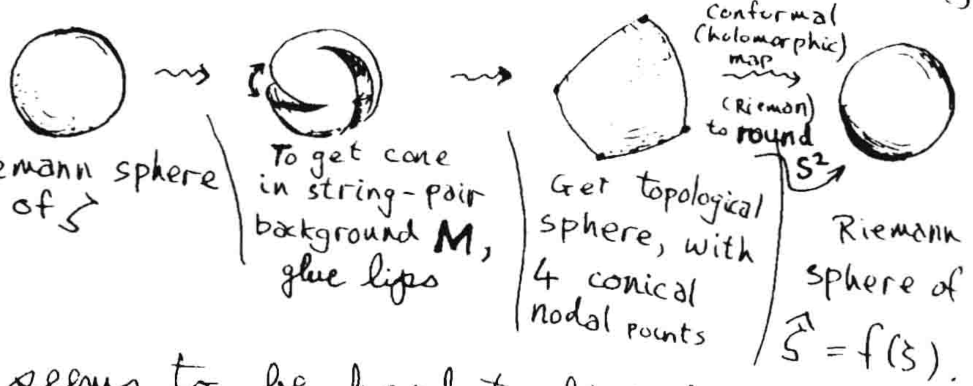


it is the other way around. (All the masses for these "particles" are inertial masses; they have zero rest-mass.)

A more involved situation is provided by a pair of strings that collide


spatial pictures
time
gravitational impulsive expanding wave

Here the function h is obtained in the following way:


Riemann sphere of ζ
To get cone in string-pair background **M**, glue lips
Get topological sphere, with 4 conical nodal points
conformal (holomorphic) map
(Riemon) to round S^2
Riemann sphere of $\hat{\zeta} = f(\zeta)$.

It seems to be hard to find f explicitly for this case, but a Riemann theorem ensures that f exists.

References

Penrose, R. (1972) The Geometry of Impulsive Gravitational Waves in General Relativity (papers in honour of J.L. Synge) (Clarendon Press, Oxford) 101–115.

Penrose, R. & MacCallum, M.A.H. (1972) Twistor theory: an approach to the quantization of fields and space-times Phys. Rept. **6C**, 241-315

Gleiser, R. & Pullin, J. (1989) Are cosmic strings stable topological defects? Class. Quantum Grav. **6**, L141–L144.

Nutku, Y. (1991) Spherical shock waves in general relativity Phys. Rev. D. **44**, 3164-3168

Yavuz Nutku &

212
KINKING AND CAUSALITY

Twistor Newsletter **34** (1992), 13–18. With A. Chamblin

206 Kinking and causality. *Twistor Newsletter* **34** (1992), 13–18. With A. Chamblin

Kinking and Causality

Andrew Chamblin and Roger Penrose

Introduction

Recently, there has been some speculation along the following lines:

Suppose M is a compact spacetime, with

$$\partial M \cong \Sigma \not\cong \emptyset$$

(Σ may be single three-manifold or the disjoint union of several). Let v be a timelike vector field with respect to the Lorentz metric, and let $\text{kink}(\partial M; v)$ denote the kinking number of v with respect to ∂M (see [1] or [2]).

Recently, there has been some suspicion that there may be a relation between the topology of ∂M, along with the value of $\text{kink}(\partial M; v)$, and the existence of closed timelike curves in M. In particular it has been conjectured that if $\partial M \cong S^3$ and $\text{kink}(\partial M; v) = 0$, then there must exist closed timelike curves in M (M assumed to be space and time orientable).

In this paper, we show that the above conjecture is false (by counterexample). In fact, we prove the more general

Proposition 1 Let Σ be any closed, orientable three-manifold, $n \in \mathbf{Z}$ an arbitrary integer. Then there exists a compact causal spacetime M with $\partial M \cong \Sigma$ and $\text{kink}(\partial M; v) = n$, where v is a timelike vector field.

Proposition 2 If M is compact and causality violating, with $\partial M \cong \Sigma \not\cong \emptyset$, then there exists a continuous deformation of the metric on M such that the new spacetime with deformed metric does not possess closed timelike curves.

(Note: Deforming the metric does not alter the kinking number).

The proofs of Propositions 1 and 2 draw on the idea of the counterexample.

Construction of Counterexample

To construct the example, consider the manifold

$$M \cong \mathbf{CP}^2 \# (S^1 \times S^3) \tag{1}$$

where # denotes "connected sum". Let $e(M)$ = "Euler number of M", then $e(M) = e(\mathbf{CP}^2) + e(S^1 \times S^3) - 2$.

Since $e(\mathbf{CP}^2) = 3$ and $e(S^1 \times S^3) = 0$ we find

$$e(M) = 1 \tag{2}$$

Define the manifold M' by

$$M' \cong M - D^4 , \tag{3}$$

where D^4 is a four-ball. Then

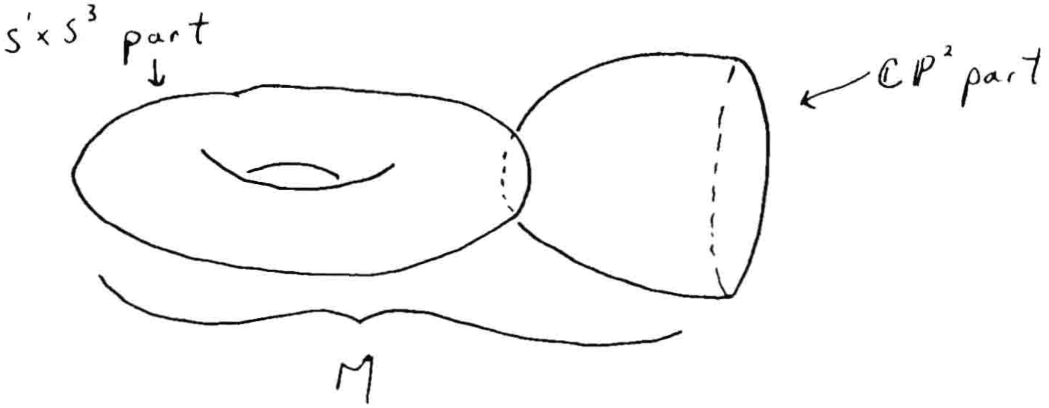
$$e(M') = 0 \tag{4}$$

Thus, we can put a nonvanishing vector field v on M' which has zero kinking on $\partial M' \cong S^3$, i.e.,

$$\text{kink}(\partial M'; v) = 0 \tag{5}$$

Now, one may suppose that there are closed timelike curves in M'; in fact, if v is outward normal on $\partial M'$ there must exist closed timelike curves (by a standard argument). However, we shall now show that we can always "cut" all of the closed timelike curves in $M' \cong M - D^4$ by choosing the D^4 that we remove from M cleverly. We shall do this "choosing" in an essentially constructive manner.

Hence, take $M \cong \mathbf{CP}^2 \# (S^1 \times S^3)$ as above and let v be a vector field on M, i.e., visually:



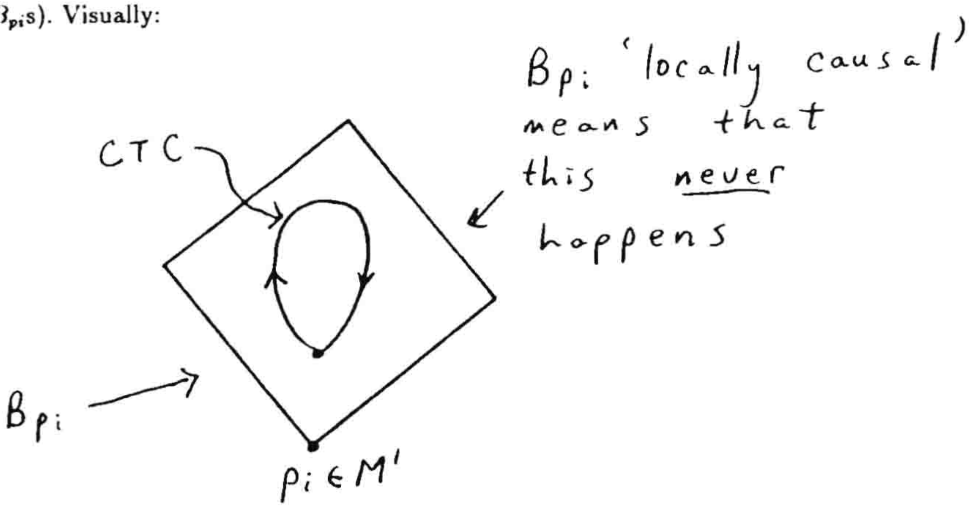
Remove a ball D^4 from around the singular point of v, so that

$$\partial D^4 \cong \partial$$

Now, we can cover M' with a finite number of sets B_{pi} of the form

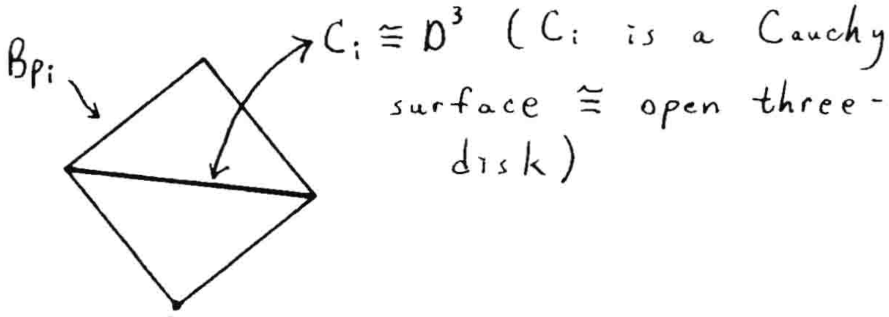
$$B_{pi} = \{x \in I^+(p_i) \cap I^-(q) \mid q \in I^+(p_i)\}$$

Furthermore, we can take the sets in this finite cover to be fine enough that they are all locally causal (i.e., no closed timelike curve, or CTC, lies entirely in any one of the B_{pi}s). Visually:

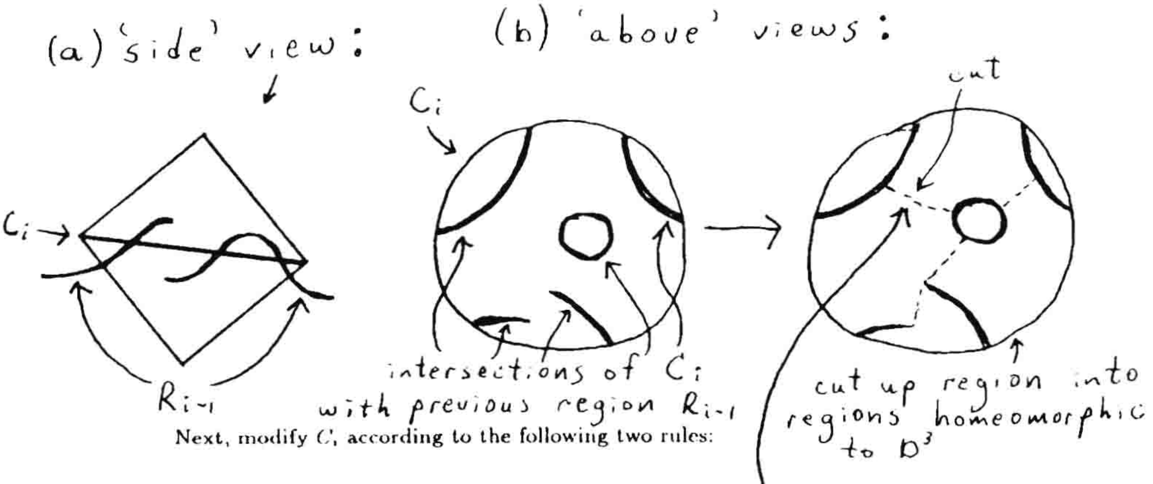

B_{pi}: 'locally causal' means that this never happens
CTC
B_{pi}
$p_i \in M'$

Now, the crucial idea of the construction depends upon our ability to cut all of the CTCs by removing a finite number of four - balls. That we can do this is reasonably intuitively obvious, but we justify this construction more rigorously as follows.

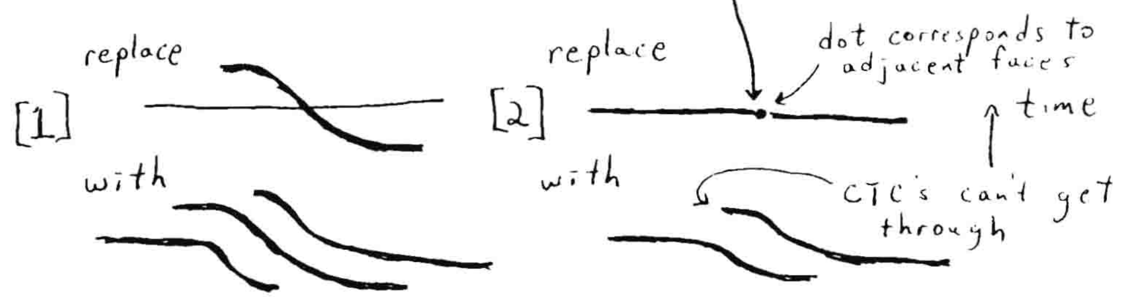
Begin by successively removing the '$t = 0$' Cauchy surface, C_i from each of our locally causal covering sets B_{pi}, as shown:


$C_i \cong D^3$ (C_i is a Cauchy surface $\cong$ open three-disk)
B_{pi}
p_i

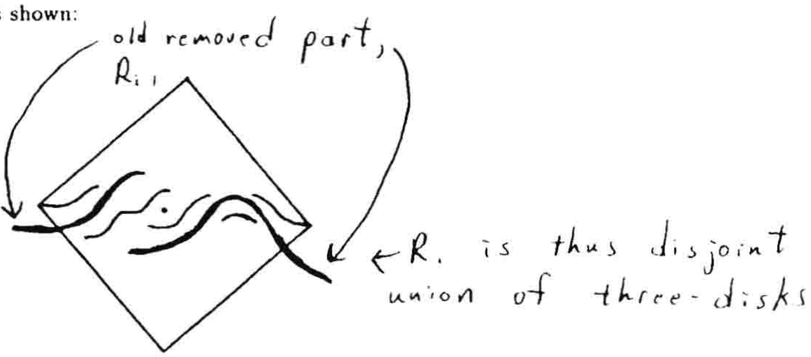
Now, at each stage C_i may already be intersected by a previously removed part (assumed to be a union of three-disks), R_{i-1}, so subdivide to get a covering of what's left by three-disks, (D^3s), as shown:


(a) 'side' view:
C_i
R_{i-1}
(b) 'above' views:
C_i
cut
intersections of C_i with previous region R_{i-1}
cut up region into regions homeomorphic to D^3

Next, modify C_i according to the following two rules:

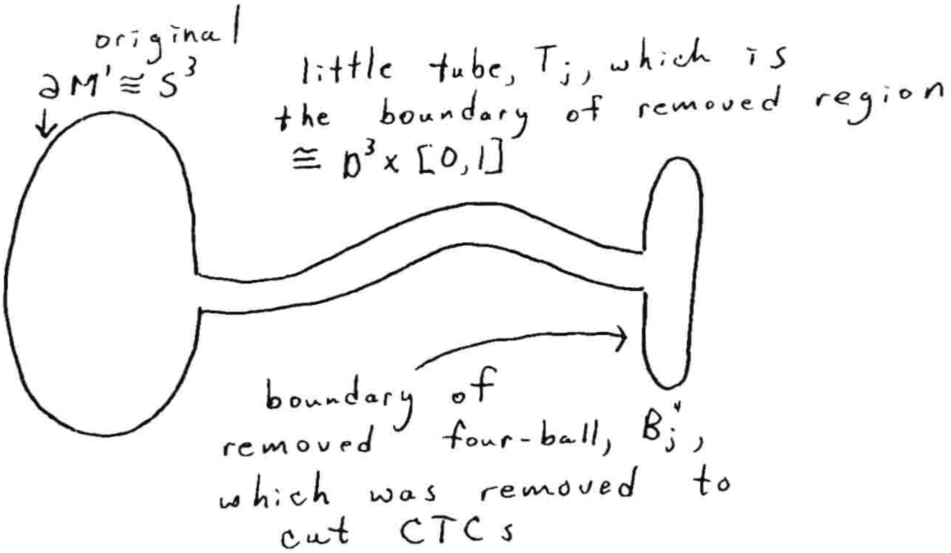

replace
[1]
with
replace
[2]
with
dot corresponds to adjacent faces
time
CTC's can't get through

Adjoin the result to R_{i-1} to get R_i, which is thus given as a disjoint union of three-balls, D^3_j, as shown:


old removed part, R_{i-1}
R_i is thus disjoint union of three-disks

Finally, thicken out the D^3_js to get disjoint four-balls B^4_js which clearly cut the CTCs.

Hence, we can cut all of the CTC's with a finite number of such four - balls. We now connect each of these 'cut out regions' to the original deleted region (i.e., where D^4 was) via 'little tubes' $T_j \cong S^3 \times [0,1]$; that is, we cut out a little tube leading from the old boundary of M' ($\partial M' \cong \partial D^4$) to the new boundary component formed by removing B^4_j, as shown:


original
∂M' ≅ S³
little tube, Tⱼ, which is
the boundary of removed region
≅ D³ x [0,1]
boundary of
removed four-ball, Bⱼ⁴,
which was removed to
cut CTCs

Call the new manifold obtained after such a finite sequence of operations 'N'. Then clearly

$$\partial N \cong S^3$$

since the total topology of the removed regions

$$R \cong D^4 \cup T_1 \cup T_2 \cup \ldots \cup T_n \cup B_1^4 \cup B_2^4 \cup \ldots \cup B_n^4$$

is still D^4, and $\partial D^4 \cong S^3$. Furthermore, v is still global and nonvanishing on N, and $e(N) = 0$; hence, $\text{kink}(\partial N; v) = 0$.

Thus, N is a causal spacetime (which is orientable) for which $\partial N \cong S^3$ and $\text{kink}(\partial N; v) = 0$; hence, N constitutes a counterexample to the conjecture mentioned in the introduction.

o o
‿

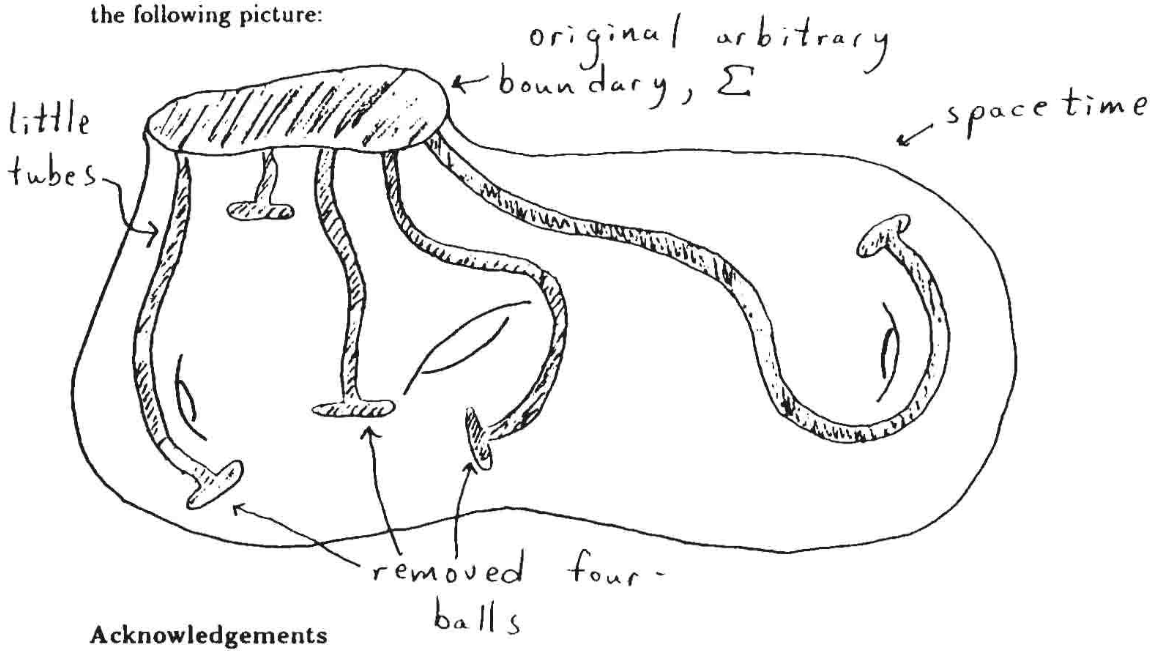
Proof of the general statement

To prove the more general Propositions (stated in the introduction) we simply generalize the above construction.

That is, let Σ be any three-manifold (or perhaps disjoint union of three-manifolds) and $n \in \mathbb{Z}$ any integer. Then we can always find a Lorentz manifold M (with timelike vector v) such that $\partial M \cong \Sigma$ and $\text{kink}(\partial M; v) = n$. This follows from the general formula

$$e(M) = \Sigma i_v \; + \; \text{kink}(\partial M; v) \tag{6}$$

(see [2]). If M should happen to possess CTC's we can always do the above construction and "cut" them by removing a finite number of four - balls B_j^4 and connecting these four - balls to the original boundary by removing little (nonintersecting) tubes T_j. (For Proposition 2, we simply continuously retract the T_js and the B_j^4s, dragging the metric with them). The fundamental idea of this paper, then, is represented in the following picture:



Acknowledgements

The authors wish thank Dr. R.P.A.C. Newman for helpful discussions. Also thanks to Jo Ashbourn for help in preparing this paper.

References

[1] G.W. Gibbons and S.W. Hawking, *Kinks and Topology Change*, DAMTP preprint

[2] H.A. Chamblin, *M.Sc. Thesis*, University of Oxford (1992), In process of completion

213

PUZZLE PAGE

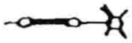
Twistor Newsletter **34** (1992). *Repeated in Twistor Newsletter* **35**

The solution to this puzzle is in Chapter 214. However, it appears that no solution to the earlier puzzle of Chapter 137 ever appeared in $\mathbb{TN}$. This is as follows. Problem B: These are axioms for what is called, I believe, a *group action* (basically a group, but where the location of the identity element has been 'forgotten'). For a group $\mathcal{G}$, whose operation is denoted by $a\circ b$, our 'abc' would stand for $a\circ b^{-1}\circ c$. To recover $\mathcal{G}$, we choose an arbitrary element of our system and call it 1 (the 'identity'). Define group multiplication by $a\circ b = a1b$ and we rapidly drive all the group axioms from the given axioms, where $a^{-1} = 1a1$. These definitions indeed yield $abc = a\circ b^{-1}\circ c$. Problem A: Since we now have all the group operations satisfied by our system, it follows that $a(bcd)e = a\circ(b\circ c^{-1}\circ d)^{-1}\circ e = a\circ(d^{-1}\circ c\circ b^{-1})\circ e = (a\circ d^{-1}\circ c)\circ b^{-1}\circ e = (adc)be$.

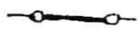
Puzzle Page (as in TN 34)

Who wrote this? When was it published?
↓ What does it mean?

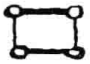
$f =$

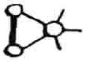
$g =$ 

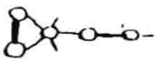
~~[illegible]~~

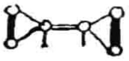
$i =$ 

~~$A = Bi$~~

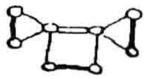
$A =$ 

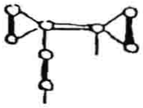
$j =$ 

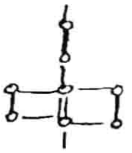
$k =$ 

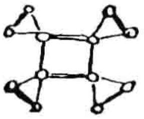
$i = \tau =$ 

$\Delta_k = A\tau - Bi = \upsilon$

$B =$ 

$\vartheta =$ 

$\upsilon =$ 

$r = C =$ 

$$D_{A\tau - Bi} = A^2C - AB^2 + AB^2$$
$$= A^2C$$

$$\Theta = \kappa^2 + \frac{A}{2}\lambda^2$$

~~$i = Bi$~~

$$J = \kappa j + \lambda k$$

$$f_F = F = \kappa f + \lambda g$$

$$g_F = \frac{1}{2}\Theta . F\Theta$$

Answers on the next page

214

SOLUTION TO PUZZLE IN TN34

Twistor Newsletter **35** (1992)

This is the solution to the puzzle of Chapter 213. Twistorians owe a great debt to William Kingdon Clifford, for reasons explained overleaf.

Solution to Puzzle in TN 34

The "twistor diagrams" were not drawn by Andrew Hodges (or by me) but by William Kingdon Clifford, the great 19th century twistor theorist (1845–1879), discoverer of Clifford algebras (on which the theory of n-dimensional spinors – and twistors – is based) and of Clifford parallels on S^3 (whose stereographic projection to Euclidean 3-space provides the twistor picture that now adorns all TN covers), and who also anticipated Einstein in suggesting that the presence of matter is to be identified with the curving of space (a profound insight that twistor theory still strives doggedly to come to terms with). After he died, at an early age (34), several of Clifford's unpublished papers were collected together posthumously and published in 1881 under the title "Mathematical Fragments, being Facsimilies of his Unfinished Papers Relating to the Theory of Graphs" — which suggests to me that those who collected the papers together did not really understand what they were about!

What are they about then? It should be borne in mind that one of the prevailing interests among algebraists of the late 19th century was the theory of invariants and, in particular, the construction of "complete sets of invariants" for binary forms. A binary form is a homogeneous polynomial (normally with complex coefficients) in two variables ξ, η. (Three or more variables was considered to be too hard, at that time.) An invariant of a binary form (or collection of such forms) would be a polynomial in the coefficients of the form(s) which is invariant under (complex) unimodular linear transformations of ξ and η. A covariant would be a corresponding thing, but which is a homogeneous polynomial in ξ, η. Nowadays, we would use tensor algebra to construct such things, but in the late 19th and early 20th century, the "symbolic calculus" was widely used. This calculus employed a strange notation that can be described roughly as follows, using a comparison with the spinor notation that is more familiar to us now.

Spinor notn.	Symbolic notn.
Take $x^0 = \xi$, $x^1 = \eta$	All factors commute,
All spinors symmetric	But $(\alpha\beta) = -(\beta\alpha)$

Examples	Spinor notn.	Symbolic notn.
forms	$\alpha_A x^A$	$\alpha_x = \beta_x = \ldots$
	$\alpha_{AB} x^A x^B$	$\alpha_x{}^2$ or $\alpha_x\alpha_x$ or $\beta_x{}^2$ etc.
	$\underbrace{\alpha_{AB\ldots L}}_{n} x^A x^B \ldots x^L$	$\alpha_x{}^n = \beta_x{}^n = \gamma_x{}^n = \ldots.$
invariants	$\alpha_{AB}\alpha^{AB} = \alpha_{AB}\alpha_{CD}\varepsilon^{AC}\varepsilon^{BD}$	$(\alpha\beta)^2 = (\alpha\beta)(\alpha\beta)$
	$\alpha_{ABC}\alpha^{ABD}\alpha_{EFD}\alpha^{EFC}$	$(\alpha\beta)^2(\alpha\gamma)(\beta\delta)(\gamma\delta)^2$
covariants	$\alpha_{AB}\alpha_{CD}x^Ax^Bx^Cx^D$	$\alpha_x{}^2\beta_x{}^2$
	$\alpha_{AB}\alpha_{CD}\alpha_{EF}x^Ax^F\varepsilon^{BC}\varepsilon^{DE}$	$\alpha_x(\alpha\beta)(\beta\gamma)\gamma_x$

A complete set of invariants [covariants] is a set from which all others can be constructed as polynomials in members of the set.

Now Clifford, with characteristic insight, not only effectively employs tensor (or spinor) ideas, but also uses a diagrammatic notation (similar to that described in the appendix to vol. 1 of Spinors and Space-Time). On the Puzzle Page, TN34, we have (all expressions up to sign — I'm not sure how Clifford dealt with signs)

f standing for, say α_{ABCDE} (i.e. for $\alpha_{\ldots}x^{\cdot}x^{\cdot}x^{\cdot}x^{\cdot}$, i.e. $\alpha_x{}^5$)
i " " $\alpha_{ABCD(E}\alpha_{F)}{}^{ABCD}$ (i.e. $\alpha_{ABCDE}\alpha_F{}^{ABCD}x^Ex^F$, i.e. $(\alpha\beta)^4\alpha_x\beta_x$)
(I'm not sure abt. symmetry bracket.)
g " " $\alpha_{ABCD(E}\alpha^{ABCDF}\alpha_{GHJK)F}$ (i.e. $(\alpha\beta)^4(\gamma\beta)\alpha_x\gamma_x{}^4$).
etc. The thick lines seem to be 4-fold lines.

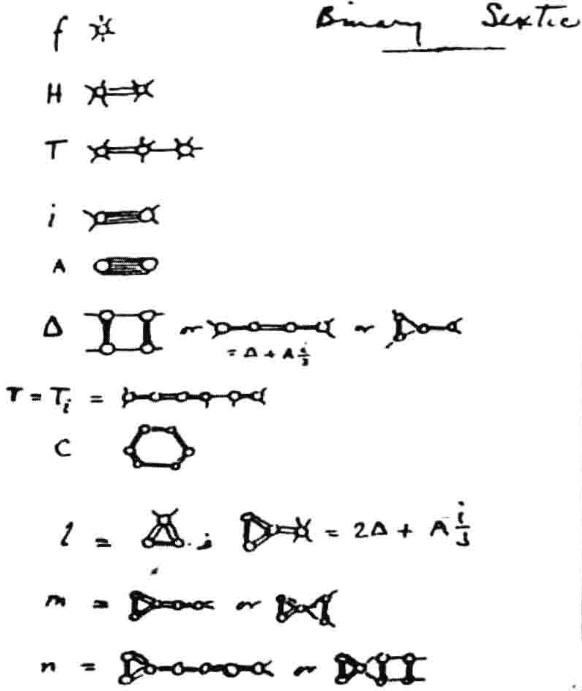
At the bottom of p. xxii Clifford seems to be working out a (complete?) set of covariants for a binary cubic (black spots) together with a binary quartic (white spots). Note the times at the left-hand side, evidently indicating how long it took him to work things out, using his notation.

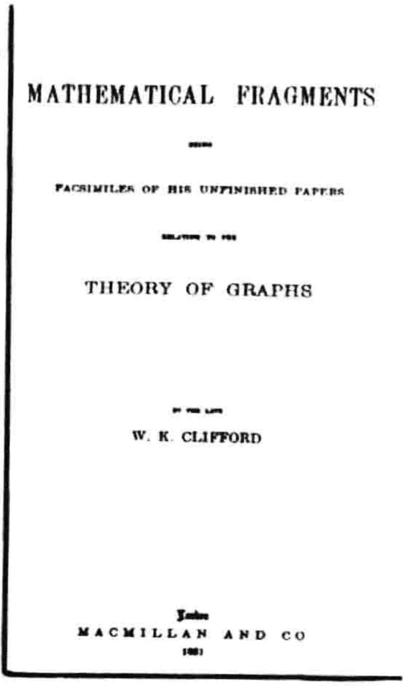
All credit to K.P.T. for spotting this publication among the books being cleared out of St Johns library.

Roger

Reference

Grace, J.H. & Young, A. (1903) The Algebra of Invariants (Camb. Univ. Press)





MATHEMATICAL FRAGMENTS

FACSIMILES OF HIS UNFINISHED PAPERS

THEORY OF GRAPHS

W. K. CLIFFORD

MACMILLAN AND CO

these are really equivalent because i is symmetrical

Table of invariants of ~~cubic &~~ Quadratic & cubic, $f_2 \cdot \varphi_3$

, 0 $D = (ab)^2$

, 4 $R = (\Delta\Delta')^2$

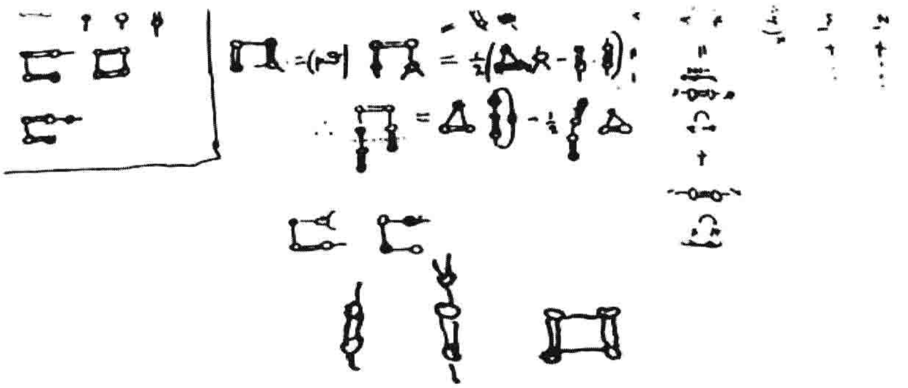
, 2 $E = (a\Delta)^2$

, 2 $F = (ap)^2 =$

$M = (ap)(ar)$

here is a series of invariants of a quadratic and any other binary form analogous to the resultant: viz. $(x\alpha . x\beta . y\alpha . y\beta)^k \left(\overline{x\alpha . y\beta}^{\,n-k} + \overline{x\beta . y\alpha}^{\,n-k}\right)$ where x, y are factors of the quadratic and $\alpha_x^n = \beta_x^n$ is the form of the n^{th} order. This invariant is the condition that one element of the quadric should be the k^{th} polar of the other.

ine in the case of a conic C_2 and a curve K_n ~~[illegible]~~ there is a curve of

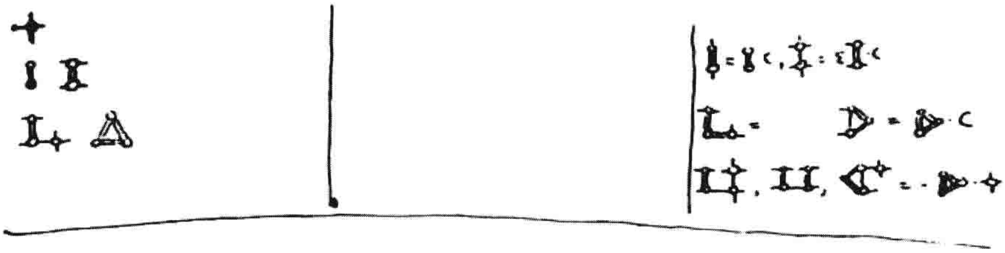


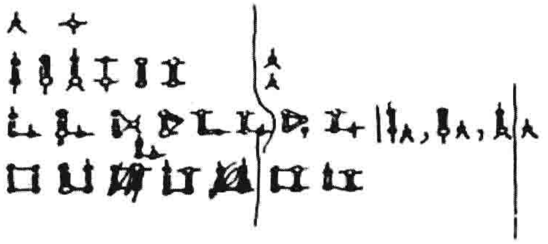
a(xu).b(yu) + a(yu). b(xu) = a(yu).b(yu).|xy|

a(xu).b(yu). c(yz) + a(yu). b(xu).c(yz) = − a(yu).b(yu).. c(xz)

a(yu). b(xu).c(yz) + a(yu) . b(xu) . c(yz) = a(yu).b(xu) c(yx).. |xz|

∴ a(xu).b(yu). c(yz) + c(xy).a(uy). b(uz)





½ hr
10 min

xxii.

215

GRAVITATIONAL COLLAPSE OF THE WAVE FUNCTION: NEW THOUGHTS

Twistor Newsletter **35** (1992)

This 𝕋ℕ article describes the change of heart I had while in Argentina shortly before my plenary address at GR13 (13th International Conference on General Relativity and Gravitation, Cordoba, Argentina, 1992). See also Chapter 218. The reasoning in both these articles is a bit muddy, but already towards the end of the present 𝕋ℕ article we find the 'Diósi–Penrose' criterion for the spontaneous wavefunction collapse time. (I was not properly aware of Diósi's earlier, but differently motivated, work at the time, but I referenced him in Chapter 218.)

Gravitational Collapse of the Wave Function: New Thoughts

A desire for a non-local basis for physical reality, as a means of making sense of the puzzling non-locality that occurs with Einstein-Podolsky-Rosen-type quantum measurements was one powerful original motivation underlying twistor theory. According to this "twistor" view, it will not be possible to provide an adequate understanding of the state-vector reduction that quantum measurement induces, in ordinary space-time terms. In TN 26, p.25, I indicated a (somewhat fanciful) picture of how a twistor-type picture might conceivably provide some kind of resolution of the non-local puzzles that state-vector reduction confronts us with, but a properly coherent viewpoint must await appropriate fundamental changes in our present theory.

Despite the fact that we are a long way from a coherent theory, I believe that it has been possible to make some new progress which is consistent with the motivational ideas that I have described elsewhere (cf. TN 22, p.1-3, and "The Emperor's New Mind" Chapter 8) — that state-vector reduction should be a gravitational phenomenon. But the new suggestions that I am presenting here differ quite significantly from the specific proposals that I made earlier.

The starting point for this development is an expression that had some relevance to my earlier ideas, namely the integral that describes the symplectic form $\{\underset{1}{K},\underset{2}{K}\}$ on the space of spin 2 massless fields in Minkowski space, whose linearized curvatures are given by

$$K_{abcd} = \Psi_{ABCD}\,\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\overline{\Psi}_{A'B'C'D'}\,.$$

This is given (up to a factor) by

$$\{\underset{1}{K},\underset{2}{K}\} = \int_{\Sigma}\left(\underset{1}{\Psi}_{ABCD}\,\underset{2}{\overline{\eta}}_{A'}{}^{BCD} + \underset{1}{\overline{\Psi}}_{A'B'C'D'}\,\underset{2}{\eta}_{A}{}^{B'C'D'}\right) d^3x^{AA'}$$

taken over a (normally spacelike) 3-surface Σ. Integrati by parts we find, schematically

$$\int \underset{1}{\Psi}\,\underset{2}{\overline{\eta}} = -\int \underset{1}{\gamma}\,\underset{2}{\overline{\chi}} = \int \underset{1}{\chi}\,\underset{2}{\overline{\gamma}} = -\int \underset{1}{\eta}\,\underset{2}{\overline{\Psi}} \quad (\text{whence } \{\underset{1}{K},\underset{2}{K}\} = -\{K\ K$$

where we have a (Dirac) chain of potentials

$$\nabla_{BB'}\eta_A{}^{B'C'D'} = \chi_{AB}{}^{C'D'}, \quad \nabla^{AA'}\eta_A{}^{B'C'D'} = 0,$$
$$\nabla_{CC'}\chi_{AB}{}^{C'D'} = \gamma_{ABC}{}^{D'}, \quad \nabla^{AA'}\chi_{AB}{}^{C'D'} = 0$$
$$\nabla_{DD'}\gamma_{ABC}{}^{D'} = \psi_{ABCD}, \quad \nabla^{AA'}\gamma_{ABC}{}^{D'} = 0$$
$$\nabla^{AA'}\psi_{ABCD} = 0$$

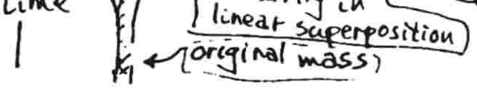
, all spinors η, χ, γ, ψ being symmetric.

N.B.
$\gamma_{...}$ are linearized spin-coefficients and $\Gamma \sim \gamma + \bar{\gamma}$
Christoffel
Also $h \sim \chi + \bar{\chi}$
lin. metric

The symplectic form $\{,\}$ is closely related to the scalar product $\langle | \rangle$, but for that a positive/negative frequency split must be made. In fact $\langle K|K\rangle = i\{K, JK\}$ where J multiplies the positive-frequency part of K by i and the negative-frequency part by $-i$.

The idea underlying my earlier proposal was that for a quantum state consisting of a superposition of two significantly different ψ_1 and ψ_2 fields, state-vector reduction would take place when the difference between these states reached roughly the (longitudinal) "one-graviton level". This might be measured in terms of the "graviton number" in the difference field $\psi_1 - \psi_2$, which might be obtained as an appropriate integral over a spacelike hypersurface. The view had been that some measure of longitudinal graviton number would accordingly be required, but this leads to profound problems, owing to the fact that the concept is not gauge invariant. The present idea is to use something closer to the symplectic form and to try to examine the Newtonian limit — since most situations of wave-function collapse occur when gravitational fields are weak and velocities small.

Consider a simple "Schrodinger's cat" type of situation in which a quantum measurement is to be achieved by moving a mass from one location to another. We thus have a superposition of the mass in one place with the mass in another. The question, according to the present point of view is: at what stage, as the masses are moved apart, should collapse occur? It appears that in the Newtonian limit, it makes more sense to do something like using a timelike hypersurface



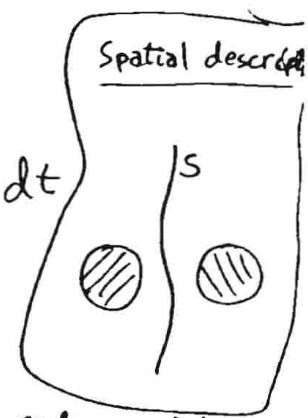
between the two masses and to perform the symplectic form integral for the two fields on that hypersurface. We take collapse to occur roughly when the region of integration is large enough that the integral reaches order unity. The integral can be written

$$\int_\Sigma \gamma_1 \bar{\chi}_2 + \int_\Sigma \bar{\gamma}_1 \chi_2 \sim \int_\Sigma (\Gamma_1 h_2 - \Gamma_2 h_1)$$

which in Newtonian terms becomes

$$\int_t^{t+T} \Big(\int_S (\phi_2 \vec{\nabla} \phi_1 - \phi_1 \vec{\nabla} \phi_2) \cdot d\vec{x} \Big) dt$$

$$= T \Big(\int_S (\quad " \quad) \cdot d\vec{x} \Big).$$



Here T is the total time that the surface S persists for, so for the integral to be order unity we require $T (= \sim \text{collapse time}) = \dfrac{1}{\int_S (\phi_2 \vec{\nabla} \phi_1 - \phi_1 \vec{\nabla} \phi_2)}$.

Note that $\phi_2 \vec{\nabla} \phi_1 - \phi_1 \vec{\nabla} \phi_2$ is divergence-free in vacuum, so the integral is unchanged as S is moved continuously without crossing matter. When matter density ρ is present we see $\vec{\nabla} \cdot (\phi_2 \vec{\nabla} \phi_1 - \phi_1 \vec{\nabla} \phi_2) = -4\pi \phi_2 \rho_1 + 4\pi \rho_2 \phi_1$, so moving S across one mass, we find that our expression is just the gravitational energy of one mass in the gravitational field of the other ($\times 4\pi$). the collapse time this suggests is the reciprocal of this energy.

As it stands, this cannot be quite right, but it suggests a much more plausible expression, namely the energy that it would take to move the two masses apart from their initial positions of coincidence (gravitational energy only). It is worthwhile to consider the Newtonian expression in the case of a uniform sphere, mass m, radius a, distance between centres b. We get, with $\lambda = b/2a$

$$\text{Energy to separate} = \begin{cases} \dfrac{m^2 G}{a} \left(6/5 - 2\lambda^2 + \tfrac{3}{2}\lambda^3 - \tfrac{1}{5}\lambda^5 \right) & (0 \le \lambda \le 1) \\ \dfrac{m^2 G}{2a\lambda} & (1 \le \lambda). \end{cases}$$

This is $\sim \dfrac{m^2 G}{a}$. In natural units ($G = c = 1 = \hbar$; gram $= 10^5$, sec $= 10^{43}$, cm $= 10$
we get: for a neutron, $T \sim 10^{60}$ = Hubble time
" " droplet of water of $a = 10^{-5}$ cm, $T = 10^{48}$ = day ; $a = 10^{-4}$ cm, $T = 10^{43}$ = s
$a = 10^{-3}$ cm, $T = 10^{38} = 10^{-5}$ sec. All this seems fairly plausible.

Thanks especially to Ted Newman and Abhay Ashtekar. ~

216

SETTING THE SCENE: THE CLAIM AND THE ISSUES

The simulation of human intelligence (Wolfson College Lectures). Ed. D. Broadbent. Blackwell, Oxford, 1992

I was somewhat surprised—but flattered—to be asked to deliver this talk initiating the series of Wolfson College Lectures, on the simulation of human intelligence. An awkward error in the wording appears eight lines up on p. [15], at the far right, where 'to be sound', should be 'does not stop'. My 'new book' referred to on p. [30], 6 lines up (in the notes) is *Shadows of the Mind* (Oxford University Press, 1994).

226 Setting the scene: The claim and the issues. *The simulation of human intelligence* (Wolfson College Lectures). Ed. D. Broadbent. 1992

Setting the Scene: the Claim and the Issues

Roger Penrose

Artificial minds?

The advent of high-speed electronic computer technology has provided a new reality, and perhaps also even an urgency, to the arguments concerning some of the most basic and long-standing of philosophical issues. If computers can, before too long, achieve a genuine artificial intelligence, then may they not also have artificial minds? We are used to the idea of artificial hips, or legs or arms, or artificial kidneys or even hearts. Why not an artificial brain? Would an artificial brain have to have a mind? Somehow, we feel that it would not be much use to us if it did not!

The whole question hangs on the resolution of deep philosophical issues. What is a mind? Are minds merely features that inevitably arise in association with the physical actions of (say human) brains? Is the essential feature of such a physical action the carrying out of an immensely complicated computation, of the particular sort that one supposes is being performed by a human brain during the act of thinking? Or is there something else that the brain does that cannot be described in merely computational terms? Perhaps minds are not even things that can be considered in scientific terms at all. Perhaps minds do not really exist.

Electronic computers offer us the possibility of objects that might ultimately be able to calculate more effectively in every way than biological brains, being not limited by biological restrictions such as imprecise structuring and the need to grow from a single cell. Already computers can be programmed to play chess better than all but a very small number of human experts. It seems not unreasonable to suppose that in many other areas where human intelligence had seemed to reign supreme the computers will overtake even the best of us. Are such 'artificially intelligent' systems *really* intelligent? Does actual intelligence require a degree of awareness and understanding of what one is doing?

In my own use of the word 'intelligence', I would take this last to be the case. I would not use the word 'intelligence' unless some actual understanding is involved; nor would I like to use the word 'understanding' unless I am prepared to believe that there is some element of awareness present. This does not remove the problem, of course. How will we be able to tell if an electronic system is 'actually' intelligent, rather than merely *simulating* intelligence, if we need to know whether it is aware or not in order to be able to tell the difference? How might we be able to tell if an electronic system is aware or not? Indeed, how can we tell whether human beings other than ourselves are aware or not? With human beings, we try to judge this from the way they talk or move, by their actions, by their writings or other creations, and so on – all *external* characteristics of their behaviour. Should we not apply similar operational criteria to electronically controlled robots? Would such a robot that *appears* to behave intelligently necessarily 'actually' possess awareness and understanding? Would it be possible to tell? Does the question have genuine meaning? While I shall certainly not be able to provide definitive answers to these age-old conundrums, I shall outline, below, various possible alternative viewpoints; and I shall also give reasons for the beliefs on this issue that I hold to myself.

The prospect of genuine artificial intelligence

While, on the whole, rather little has been achieved by artificial intelligence to date – in any important area of intellectual expertise where genuine human understanding and insight has seemed to be important – there have been some very impressive developments in recent years, most particularly the development of powerful chess computers. But for the most part, the very optimistic claims that had sometimes been made by proponents of artificial intelligence and promoters of expert systems, etc., have not yet been fulfilled. But these are still very early days, if we are to consider what artificial intelligence might ultimately achieve. There would appear to be some enormous potential advantages over brains that electronic computers could eventually make use of. Electronic circuits are already about a million times faster that the firing of the neurons in the brain. Furthermore, they have an immense precision in timing and accuracy of action that is in no way shared by our own neurons. There is a great deal of randomness in the brain's 'wiring' that could apparently be vastly improved upon by the deliberate and precise organization of electronic printed circuits. Moreover, in those places where the brain now does have an advantage, it would seem that this advantage may be short-lived. In its total neuron number – some hundreds of thousands of millions – the brain is perhaps temporarily ahead (although this depends, to some extent, upon how one does that calculation, for example because different computer units can always be added together to form larger and larger ones). Also, there are, on the average, a good deal more connections between different neurons than there are connections between transistors in a computer. But it is clear, especially when one takes into account the rapid increase in computer technology over the years, that these few remaining *numerical* advantages that the brain may have will not last longer than a few decades at the very most. Moreover, there may well be technological revolutions waiting in the wings, such as the replacing of the wires and transistors of our present computers by appropriate optical (laser) devices, thereby achieving enormous increases in speed, power, and miniaturization. There seems to be little doubt that on any issue of merely computing power, if computers do not have the advantage over brains already, then they *will* certainly have it before too long.

But are the relevant issues merely those of computing power, or speed, or memory, or perhaps of the detailed way in which things are 'wired up'? Might we, on the other hand, be doing something with our brains that cannot be described in computational terms at all? How do our feelings of conscious awareness – of happiness, pain, love, aesthetic sensibility, will, understanding, etc. – fit into such a computational picture? Will the computers of the future indeed actually have *minds*? Does it make sense to talk about such things in scientific terms at all or is science in no way competent to address issues that relate to the human mind? It seems to me that there are at least four different viewpoints[1] (or extremes of viewpoint) that one may reasonably hold on the matter:

1 All thinking is computation, and the mere carrying out of appropriate computations will evoke feelings of conscious awareness.
2 It is the brain's physical action that evokes awareness and any physical action can in principle be simulated computationally, but computational simulation by itself cannot evoke awareness.
3 Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
4 Awareness cannot be explained in physical, computational, or any other scientific terms.

The point of view expressed in **4**, which negates the physicalist position altogether and regards the mind as something that is entirely inexplicable in scientific terms, is the viewpoint of the mystic; and at least some ingredient of **4** seems to be involved in the acceptance of religious doctrine. My own position is that questions of mind, though they lie uncomfortably with present-day scientific understanding, should not be regarded as being forever outside the realms of science. If science is as yet incapable of saying much that is of significance concerning matters of the mind, then science itself will have to change. While I reject mysticism in its negation of scientific criteria for the furtherance of knowledge, I believe that within an expanded science and mathematics there will be found sufficient scope to accommodate even what is needed for an understanding of mind. I shall expand on some of these ideas later on, but for the moment it will be sufficient to say that under my own viewpoint, I am rejecting **4**; and I am attempting to move forward along the path that science has set out for us. I believe that this route will eventually reveal enough mysteries of its own, so that it may not seem so strange if the mystery of mind may be ultimately explicable in terms of it.

Let us consider what seems to be the opposite extreme to **4**: the viewpoint **1**. This is what is referred to as *strong AI* (strong artificial intelligence) or *functionalism.*[2] It is regarded by some as the only viewpoint that an entirely scientific attitude allows. Others would take **1** to be an absurdity that is barely worth serious attention. It is the viewpoint that the arguments of my book, *The Emperor's New Mind* (Penrose, 1989) were most specifically directed against. The length of that book alone (466 pages in hardback) should make it clear that, while I do not myself believe that **1** is correct, I do regard it as a serious possibility that is worthy of considerable attention. It is an implication of a highly operational attitude to science, where, also, the physical world is taken to operate entirely computationally. In one extreme of this view, the universe itself is taken to be, in effect, a gigantic computer,[3] and appropriate sub-calculations that this computer performs will evoke the feelings of awareness that constitute our conscious minds. I suppose that this viewpoint – that physical systems are to be regarded as merely computational entities – stems partly from the belief that physical objects are themselves merely 'patterns of information', in some sense.

Even if we do not think that it is appropriate to regard the universe as simply being a computer, we may feel ourselves operationally driven to viewpoint **1**. Suppose that we have a robot that is controlled by a computer and which responds to questioning exactly as a human would. We ask it how it feels, and find that it answers in a way that is entirely consistent with its actually possessing feelings. It tells us that it is aware, that it is happy or sad, that it can perceive the colour red, and that it worries about questions of 'mind' and 'self'. It may even give expression to a puzzlement about whether or not it should accept that *other* beings (especially human beings) are to be regarded as possessing a consciousness similar to the one that it feels itself. Why should we disbelieve its mere *claims* to be aware, to wonder, to be joyful, or to feel pain, when it might seem that we have as little to go on with respect to other human beings whom we *do* accept as being conscious? The operational argument does, it seems to me, have some considerable force, even if it is not entirely conclusive. If all the *external* manifestations of a conscious brain can indeed be simulated entirely computationally then there would be a case for accepting that its *internal* manifestations – consciousness itself – are also present in association with such a simulation.

What about **2**? I think that it is the viewpoint that many would regard as the 'common-sense' one.[4] Like **1**, it affirms a view that all the physical objects of this world must behave according to a science that, in principle, allows that they can be computationally simulated. On the other hand, it strongly denies the operational claim that a thing that behaves externally as a conscious being must necessarily be conscious itself. As the philosopher John Searle has stressed, a computational simulation of a physical process is indeed different from that process itself.[5] (A computer simulation of a hurricane, for example, is certainly no hurricane!) On view **2**, the presence or absence of consciousness would depend very much on what actual physical object is carrying out a computation. Thus, the action of a biological brain might evoke consciousness, while its electronic simulation might not. It is not necessary, in viewpoint **2**, for this distinction to be between biology and physics. But the actual *physical* constitution of the object in question (say a brain), and not just its computational action, is regarded as all-important.

The viewpoint **3** is the one which I believe myself to be closest to the truth. It is more of an operational viewpoint than **2** since it asserts that there are external manifestations of conscious objects (say brains) that differ from the external manifestations of a computer: consciousness cannot even be properly *simulated* computationally. I shall be giving my reasons for this belief in due course. Since **3**, like **2**, maintains the physicalist standpoint that minds arise as manifestations of the behaviour of certain physical objects (brains – although not necessarily only brains), it follows that an implication of **3** is that *not* all physical action can be properly simulated computationally. It is not completely clear whether present-day physics allows for the possibility of an action that is in principle impossible to simulate on a computer. Rather little is known of a precise mathematical nature on this issue. My own opinion is that such non-computational action would have to be found in an area of physics that lies outside the presently known physical laws. Later on, I shall briefly indicate some of the reasons, coming from within physics itself, for believing that a new understanding is indeed needed, in an area that lies intermediate between the 'quantum level' of molecules, atoms and subatomic particles and the 'classical level' of everyday objects.

However, it is not by any means universally accepted, among physicists, that such a new physical theory is required.

What of the future?

What do the viewpoints **1**, **2**, **3**, **4** tell us to expect for the future of this planet? According to 1, there will come a stage when appropriately programmed supercomputers will reach – and then race beyond – all human mental capabilities. Of course different people who hold to **1** could have very different views as to the time-scales involved in this. Some might reasonably take the line that it will be many centuries before computers will reach our level, so little being presently understood about the calculations that the brain must indeed be performing (they would claim) in order to achieve the subtlety of action that we undoubtedly attain – a subtlety that would be necessary before appreciable 'awareness' would take place. Others argue for a much shorter time-scale. In particular, Hans Moravec, in his book *Mind Children* (1988), makes a reasoned case in support of his claim that 'human equivalence' will be reached in a mere forty years, basing his arguments on the rate at which computer technology has moved forward at an accelerating rate over the past half century, and on the proportion of the brain's activity that has already been successfully simulated. (Some have argued for a much shorter time-scale, sometimes, even, where the predicted date for human equivalence has already passed!) Lest the reader feel dismayed by the prospect of being overtaken by computers in (say) forty years' time, hope is offered – indeed promised – by the assured prospect of our being able to transfer our 'mind programs' into the shining metallic bodies of the robots of our choice, thereby obtaining for ourselves a form of immortality (Moravec, 1988).

Such optimism is not available to the holders of viewpoint **2**, however. Their standpoint does not differ from **1** with regard to what computers will ultimately be able to achieve. Whether it would take centuries or a mere forty years, computers would again be expected to reach human equivalence, and then race beyond whatever we are able to achieve with our relatively puny brains. Now the option is not open to us to 'join' the computer-controlled robots, and it would seem that we must resign ourselves to the prospect of a planet ultimately controlled by insentient machines! Of all the viewpoints **1**, **2**, **3**, **4** it seems to me that it is **2** that offers the most pessimistic view of the future of our planet – despite its apparently 'common-sense' nature!

According to **3** or **4**, on the other hand, it would be expected that the computers, as we understand them today, would (or should) always remain subservient to us, no matter how far they advance with respect to speed, capacity and logical design. Viewpoint **3**, however, is open with regard to future scientific developments that might lead to the construction of devices – *not* based on today's concept of computers – which could achieve *actual* intelligence and awareness. Since we at present lack almost all the scientific understanding that would be necessary, let alone any of the technological know-how, little can be gained from such speculation at the present time.

Irrespective of such digressions into futurology, we must ask: what is it that we actually *do* with our brains? Is 'computing', in one form or another, the only possibility? Might we, on the other hand, sometimes be doing something *else* with our brains and minds – something very different from computation, for which our very awareness and (apparent?) free will may be playing an essential role? Are these things that can be discussed at all in anything like scientific terms – as would be denied by viewpoint **4**? I believe that they can be, and that viewpoint **3** is a genuine scientific possibility, although we must be prepared for the eventuality that our scientific criteria and methods may undergo subtle but important shifts. We must be prepared to examine clues that may present themselves in unexpected ways, in areas of genuine understanding that may at first appear to be largely irrelevant.

The non-computability of mathematical thinking

In a moment, we shall need to turn to some mathematics. It is there that our thinking processes have their purest form. If thinking is just carrying out a computation of some kind, then, we ought to be able to see this most clearly in our mathematical thinking. Yet, remarkably, the very reverse turns out to be the case. It is within mathematics that we find the clearest evidence that there must actually be something in our conscious thought processes that eludes computation.

Nevertheless, one might reasonably ask why we shall need to turn to mathematics at all, in order to find a suitable area in which to demonstrate that there must be something non-computational in conscious thought processes. Indeed, mathematics is certainly very far from being the *only* animal activity requiring consciousness. It is a very specialized human activity, whereas the phenomenon of consciousness is commonplace, being likely to be present in much non-human as well as human mental activity. The reason for needing to address the question of consciousness here in a mathematical context is that it is only within mathematics that we can expect to find anything approaching a rigorous demonstration that some, at least, of conscious activity *must* be non-computational in nature. (The issue of computation, by its very nature, is after all a mathematical one.) If it can be demonstrated that whatever we do with our brains or minds when we consciously understand *mathematics* is different from anything that we can achieve by use of a computer, then the possibility of non-computational thinking generally has been established. Moreover, since understanding seems to be a conscious activity, it is *conscious* thinking that seems to have this non-computational character.

Of course consciousness has many other manifestations. The perception of the colour red, for example, is something that requires consciousness, as is the sensation of pain, the appreciation of a melody, the bringing to mind of an early memory, the willed action to get up from one's bed, and so on. Some might dispute an all-embracing role of some single concept of consciousness in all of these various manifestations. In my own view, there is indeed a unified concept of 'consciousness' that is central to all these separate aspects of mentality, but it will not be necessary to worry about this here. Let us just concentrate on 'understanding', and this will include 'mathematical understanding'.

In order to address this issue, we shall be concerned here with *calculations*. By a calculation I mean, in effect, the action of a general purpose computer operating according to some computer program. (Strictly speaking, the computer must be idealized in that it has an in principle unlimited storage capacity – which could be in the form of an unlimited supply of floppy discs, say. The appropriate idealized mathematical concept is that of a *Turing machine*, but it will not be necessary to go into the details of what this term means here.) Another word for a calculation is an *algorithm*. Operations that can be achieved (in principle) calculationally are referred to as *algorithmic*. It should be realized that calculations (algorithms) are not merely the performing of ordinary operations of arithmetic, such as just adding or multiplying numbers together, but can involve other things also. Well-defined *logical operations* can also be part of a calculation. For an example of a calculation, we might consider the following task:

Find a number that is not the sum of three square numbers. **(A)**

By 'number', I mean here a 'natural number', i.e. one of

0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...

A *square* number is the product of a natural number by itself, i.e. one of

$0 = 0^2$, $1 = 1^2$, $4 = 2^2$, $9 = 3^2$, $16 = 4^2$, $25 = 5^2$,
$36 = 6^2$, ...

The calculation (A) could then proceed as follows. We try each natural number in turn, starting with 0, to see whether or not it is the sum of three squares. We need only consider squares that are no larger than the number itself. Thus, for each natural number, there are only finitely many square numbers to try. As soon as three square numbers are found that do add to it, then our calculation moves on to the next natural number, and we try again to find a triplet of squares (less than the number) that sum to it. Our calculation stops only when a natural number is found for which each such triplet of squares fails to add to it. To see how this works, start with 0. This is $0^2 + 0^2 + 0^2$ so it is indeed the sum of three squares. Next we try 1 and we find that although it is not $0^2 + 0^2 + 0^2$ it is indeed $0^2 + 0^2 + 1^2$. Our calculation tells us now to move on to 2 and we ascertain that although it is not $0^2 + 0^2 +$

0^2 or $0^2 + 0^2 + 1^2$, it is indeed $0^2 + 1^2 + 1^2$; we then move on to 3 and find $3 = 1^2 + 1^2 + 1^2$; then to 4, finding $4 = 0^2 + 0^2 + 2^2$; then $5 = 0^2 + 1^2 + 2^2$; then after finding $6 = 1^2 + 1^2 + 2^2$ we move on to 7, but now all triplets of squares (each member no greater than 7)

$0^2 + 0^2 + 0^2$. $0^2 + 0^2 + 1^2$. $0^2 + 0^2 + 2^2$. $0^2 + 1^2 + 1^2$.
$0^2 + 1^2 + 2^2$. $0^2 + 2^2 + 2^2$. $1^2 + 1^2 + 1^2$. $1^2 + 1^2 + 2^2$.
$1^2 + 2^2 + 2^2$. $2^2 + 2^2 + 2^2$.

fail to sum to 7, so the calculation halts and we reach our conclusion: 7 is a number of the kind we seek, being *not* the sum of three squares.

However, with the calculation (A) we were fortunate. Suppose we had tried, instead, the calculation:

Find a number that is not the sum of four square numbers. (B)

Now when we reach 7 we find that it *is* the sum of *four* squares: $7 = 1^2 + 1^2 + 1^2 + 2^2$, so we must move on to 8, finding $8 = 0^2 + 0^2 + 2^2 + 2^2$, etc. The calculation goes on and on: $23 = 1^2 + 2^2 + 3^2 + 3^2$, $24 = 0^2 + 2^2 + 2^2 + 4^2$, and it never seems to stop at all. In fact it never does. It is a famous theorem first proved by the great eighteenth-century mathematician Joseph L. Lagrange that *every* number is, indeed, the sum of four squares. It is not such an easy theorem (and even Lagrange's contemporary the great Swiss mathematician Leonhard Euler, a man of astounding mathematical insight, originality and productiveness, had tried but failed to find a proof). I am certainly not going to go into Lagrange's argument here, so let us instead try something very much simpler:

Find an odd number that is the sum of two even numbers. (C)

I hope that it is obvious that *this* calculation will never come to an end!

I have given two examples ((B) and (C)) of calculations that never terminate. In the first case this fact, though true, is not at all easy to ascertain, while in the second, its non-termination is obvious. Calculations may or may not terminate and, moreover, in the cases when they do not terminate it may be hard to see that they do not, or it may be very easy. We may reasonably ask: by what procedures do mathematicians convince themselves and others that certain calculations do not in fact terminate? Are they themselves following some calculational (or algorithmic) procedure in order to ascertain things of this kind?

The answer appears to be 'No'. This is essentially one of the implications of Gödel's famous incompleteness theorem, and it is important that we try to understand it. This theorem tells us that no set of rules whatsoever will be sufficient to ascertain correctly, in all cases, that non-terminating calculations do not in fact terminate. Here, by a 'set of rules' I mean some system of formalized procedures for which it is possible to check entirely computationally whether or not the rules have been correctly applied. But Gödel's theorem appears to tell us more than this, namely that the insights that are available to human mathematicians – indeed, to anyone who can think logically with understanding and imagination – cannot be completely formalized as such a set of rules. Rules can sometimes be a partial substitute for understanding, but they can never replace it entirely.

In order to see how Gödel's theorem (in the simplified form that I shall give, which depends on some ideas due to Alan Turing) demonstrates this, we shall need a slight generalization of the kind of statement about calculations that I have been considering. Instead of asking whether or not a single calculation, such as (A), (B) or (C) ever terminates, we shall need to consider a calculation that depends on – or *acts* upon – a *natural number n*. Thus, if we call such a calculation C(n), we can think of this as providing us with a *family* of calculations, where there is a separate calculation for each natural number 0, 1, 2, 3, 4, . . . , namely the calculation C(0), C(1), C(2), C(3), C(4), . . . , respectively, and where the way in which the calculation depends upon n is itself entirely calculational. This means that C(n) is the action of some algorithm applied to the number n. Think of an ordinary general purpose computer, and regard n as merely providing the 'data' for the action of some computer. What we are interested in is whether this computer action ever stops or not, for each choice of n.

In order to clarify what is meant here, let us consider two examples, slightly generalizing the ones given above:

Find a number that is not the sum of n square numbers (D)

and

Find an odd number that is the sum of n even numbers. (E)

It should be clear from what has been said above that the calculation (D) will stop *only* when $n = 0, 1, 2$ and 3 (finding the numbers 1, 2, 3 and 7, respectively, in these cases), and that (E) stops for no value of n whatever. If we are actually to ascertain that (D) does not stop when n is 4 or larger we require some formidable mathematics (Lagrange's proof); on the other hand, the fact that (E) does not stop for any n is obvious. What are the procedures that are available to mathematicians for ascertaining the non-stopping nature of such calculations generally? Are these very procedures things that can be put into a calculational form?

Suppose, then, that we have some calculational procedure A which, when it successfully terminates, provides us with a demonstration that a calculation such as $C(n)$ actually does *not* ever stop. I am certainly not requiring that A can *always* decide that $C(n)$ does not stop when in fact it does not, but I *do* insist that A does not ever give us wrong answers, i.e. that if it comes to the conclusion that $C(n)$ does not stop, then in fact it does not. (If A did give us wrong answers of this kind, then in principle we could check up on this fact because its error would be recognized as soon as $C(n)$ comes to a halt.) If A does not in fact give us wrong answers, we say that A is *sound*.

In order to understand how A could itself act as a calculation, we need to have a way of coding the different calculations $C(n)$ on which A acts. All possible different Cs can in fact be listed, say as

$$C_0(n),\ C_1(n),\ C_2(n),\ C_3(n),\ C_4(n),\ C_5(n),\ \ldots\ldots,$$

where we can take this ordering as being given, say, as some kind of numerical ordering of computer programs. (To be explicit, we could take this ordering as being provided by the Turing machine numbering given in *The Emperor's New Mind*, so that then the calculation $C_q(n)$ is the action of the qth Turing machine T_q acting on n.) One technical thing that is important here is that this listing is *computable*, i.e. there is a single calculation C that gives us C_q when it is presented with q, or, more precisely, the calculation C acts on the *pair* of numbers q, n (i.e. q followed by n) to give $C_q(n)$.

The procedure A can now be thought of as a particular calculation that, when presented with the pair of numbers q, n, tries to ascertain that the calculation $C_q(n)$ will never ultimately halt. We are going to try to imagine that A might be a formalization of *all* the procedures that are available to human mathematicians for validly deciding that calculations never will halt. But it is not necessary for us to think of A in this way just now. A is just *any* sound set of calculational rules for ascertaining that some calculations $C_q(n)$ do not ever halt. Being dependent upon the two numbers q and n, the calculation that A performs can be written $A(q,n)$, and we have:

If $A(q,n)$ stops, then $C_q(n)$ doesn't stop. (F)

Perhaps you are worrying that A might be able to decide that $C_q(n)$ actually *does* stop and would itself stop as soon as it had made that decision. But that is not what we are asking for A to do. If, in some way, our procedure comes to the conclusion that $C_q(n)$ actually does stop, then we must ensure that this conclusion makes A go into a 'loop', rather than output some finite conclusion, so that A does *not* in fact stop in the circumstance that $C_q(n)$ does stop. This is just a technical point, and it does not imply any restriction on what procedures we are considering for ascertaining that calculations do *not* stop.

Now let us consider the particular statements (F) for which q is put equal to n. This may seem an odd thing to do, but it is perfectly legitimate. With q equal to n, we now have:

If $A(n,n)$ stops, then $C_n(n)$ doesn't stop. (G)

We now notice that $A(n,n)$ depends upon just *one* number n, not two, so it must be one of the calculations C_0, (n), C_1, (n), C_2, (n), C_3, (n), . . . (as applied to a general value of n). Let us suppose that it is in fact $C_k(n)$:

$A(n,n) = C_k(n)$. (H)

Now put $n = k$. We have, from (H),

$$A(k,k) = C_k(k) \tag{I}$$

and, from (G), with $n = k$,

$$\text{If } A(k,k) \text{ stops, then } C_k(k) \text{ doesn't stop.} \tag{J}$$

Let us try to see whether the calculation $A(k,k)$ actually stops or not. First, suppose that it does. Then (J) tells us that $C_k(k)$ does not stop. But by (I), $C_k(k)$ is the same as $A(k,k)$, so it follows that $A(k,k)$ does not stop after all. This must therefore be the right answer: $A(k,k)$ does not stop. Moreover $C_k(k)$, being the same as $A(k,k)$ does not stop either. We conclude that our procedure A is incapable of ascertaining that this particular calculation $C_k(k)$ does not stop even though it does not!

At this point it might be advisable to go over the whole argument again, just to make sure that I have not 'put one over' on the reader! Admittedly there is an air of the conjuring trick about the argument, but it is perfectly legitimate, and it only gains in strength the more minutely it is examined. We have found a calculation $C_k(k)$ that we know does not stop; yet the given calculational procedure A is not powerful enough to ascertain that fact. This is the Gödel(-Turing) theorem in the form that I require. It applies to any calculational procedure A whatever, that we know to be sound, for ascertaining that calculations do not stop. We deduce that no sound set of calculational rules (such as A) can ever suffice for ascertaining that calculations do not stop. Moreover, since from the knowledge of A and its soundness, we can actually construct a calculation $C_k(k)$ that we *know* to be sound, we deduce that A *cannot* be a formalization of the procedures available to mathematicians for ascertaining that calculations do not stop, no matter what A is. Hence:

> Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth.

It seems to me that this conclusion is inescapable^6. However, many people have trouble coming to terms with it, and certainly many have argued against the stronger deduction that there must be something essentially non-calculational in our thought processes. Indeed, there are various possible loopholes to consider. Let us try to go carefully through these in turn.

Some possible viewpoints

Let us accept, on the basis of the preceding argument, that human mathematicians, when they form their views of the truth or otherwise of mathematical statements, are not using any knowably sound algorithm. What procedures might they in fact be using? Let me list a number of possibilities. They might use:

1 A horrendously complicated *unknowable* algorithm X;
2 An *unsound* (but presumably approximately sound) algorithm Y;
3 An ever-changing algorithm;
4 The environment;
5 Random elements;
6 Mysticism;
7 Non-algorithmic physics.

Let us consider these possibilities in turn.

1 Do we use a horrendously complicated unknowable algorithm X?

An important thing to realize about the insights that allow mathematicians to come to their conclusions about the truth, or otherwise, of mathematical statements is that these insights are *communicable*. Thus, if a certain argument convinces one mathematician, then the same argument can also be used to convince any other mathematician – at least in principle, but with (admittedly) considerably varying amounts of difficulty. It should always be possible to break down a mathematical argument into steps that themselves represent things that are 'obvious' to all, rather than being things that are personal to particular individual mathematicians. The same comment would apply to the 'Gödelian insight' that allows one to deduce from a belief in the soundness of a procedure for ascertaining that certain calculations do not stop (such as A above) that a particular calculation (such as $C_k(k)$ above) does not in fact stop. It is not that we are talking about different possible horrendous algorithms that might happen to be running around in different mathematicians' heads, depending upon how each one's brain might happen to be 'wired up'. We are talking about *one* horrendously complicated unknowable algorithm X that applies universally to the mathematical community as a whole.

It is true that a good deal of mathematics is often carried out within what are called 'formal systems', and these are essentially algorithmic structures, but these cannot encompass the *entirety* of the insights that are available to mathematicians, as Gödel's procedure demonstrates. Such formal systems are never so complicated that they are actually 'unknowable', but some people might consider that, in certain cases, their meanings are so esoteric that their *soundness* is unknowable. However, any system whose soundness cannot be trusted cannot itself be used without qualification in a mathematical argument. An argument based on a questionable formal system F would have to run something like 'Assuming F, then it follows that . . .' if it is to be acceptable to the mathematical community as a whole. I do not see how any such F could possibly encompass precisely the insights that are available to the mathematical community. If F is accepted as sound, then the Gödel procedure yields a mathematical proposition G that must also be accepted as sound, yet which is inaccessible by the methods of F. (Here, F plays the role of A, and G is essentially the assertion that $C_k(k)$ does not terminate.) If F is not accepted as sound, then we cannot use it (without qualification) for ascertaining the truth of mathematical propositions. Either way, F cannot precisely encompass all the insights available to human mathematicians.

It seems to me that a similar kind of argument would apply to our putative *unknowable* algorithm X. Can it be that, unknown to all of us, when mathematicians come to their conclusions about the truth of mathematical statements they are just acting according to the horrendous X? This seems to be totally at variance with what mathematicians seem *actually* to be doing when they express their arguments in terms that can (at least in principle) be broken down into assertions that are 'obvious', and agreed by all. I would regard it as far-fetched in the extreme to believe that it is *really* the horrendous unknowable X, rather than these simple and obvious ingredients, that lies lurking behind all of our mathematical understanding.

There is another argument that, in my view, adds convincing support to the belief that mathematicians do not actually use X. For how is it that X could have arisen by natural selection? One need but glance at the contents of any respectable mathematical research journal to see how far removed from everyday experience are the subject-matter and lines of reasoning that are used by mathematicians. If X is needed for doing all this, and if X is shared among all mathematicians – and, in principle, among all thinking people – then somehow X would have to have arisen by natural selection, the process that, over thousands of centuries, has made Man what he is. Yet the selective value of a potential ability to do such sophisticated mathematics must have accounted for *nil* throughout Man's history! Even today, there is very little direct survival value related to an ability to reason accurately with obscurely defined infinite sets. If such an ability depended upon the possession, within the heads of at least a significant proportion of us, of the horrendously complicated unknowable X, then one is faced with the apparently insuperable problem of comprehending how anything like X could have arisen by natural selection.

In my own view, what underlies an ability to do mathematics, whether sophisticated or not, is something very different from the putative algorithm X; it is our general (non-algorithmic) ability to *understand*. What natural selection has provided is an ability to understand one's surroundings and to reason about the implications of one's actions. Accordingly, Man's superb ability to survive must, in large measure, have arisen from his ability to understand. This ability would have enabled him to achieve an effective superiority over his less comprehending adversaries. It would have had a profound value to him for his survival in many different ways, such as in the construction of mammoth traps and the like. The quality of understanding is, perhaps fortuitously, also the crucial quality that is needed for doing mathematics. It was not an *algorithm* X that was favoured, in Man (at least) by natural selection, but this wonderful ability to understand! The Gödel argument shows that *mathematical* understanding, at least, is not an algorithmic activity, and it seems to me that we must accept that a non-algorithmic quality, namely an ability to understand, can indeed arise by natural selection.

2 *Do we use an unsound (approximately sound) algorithm Y?*

What about the possibility that mathematicians might use an algorithm Y that is only approximately sound? Mathematicians, after all, do make mistakes – not infrequently, in fact (and I write from personal experience here). It would appear that Alan Turing himself may have believed that this was the loophole in the Gödelian argument, when it is applied to the algorithmic action that he himself apparently believed applied to the actual processes taking place in the human brain. For myself, I find this possibility to be unlikely as the real explanation. Many of the above objections to X, it seems to me, would apply also to Y. Moreover, although mathematicians do indeed not uncommonly make mistakes in their personal workings and in their guesses as to the truth, it is rare for these errors to survive significantly in their published writings. The essential point is that the errors are *recognizable* as errors. When a mistake has been pointed out, either by someone else or else by the same mathematician at a later time, it is *seen* to be an error – or at least if, for some reason a mental blockage persists, such a blockage could in principle be removed by further explanation and insight. The way that mathematics progresses is not at all like the blind following of an unsound algorithm Y.

Moreover, unsound algorithms are liable to land one in contradictions. In particular, in the arguments given above, if the algorithm A were unsound, then it would incorrectly assert that some calculation $C_q(n)$ does not stop when indeed it does. If the calculation $C_q(n)$ is actually stumbled upon and performed, and is seen to come to an end while $A(q,n)$ successfully terminates, then we would know that something is amiss. Of course things like this do occur with working mathematicians: errors in an argument may be spotted when a counter-example to the conclusion is found.

But that would seem to be something very different from what is being envisaged here. Actual errors in a mathematical argument could, *in principle*, have been spotted in any case. The counterexample merely provides a route to finding something that could also be found directly, namely by checking through all the details of the original argument. In the case of Y, this putative algorithm is supposed to represent the *totality* of all the means that are available to human mathematicians. Thus Y is in principle *not* correctable by human reasoning, since whatever means are available for achieving this should already have been included in the procedures of Y. (I shall return to the issue of possible self-improving algorithms under the next heading.)

There is a type of procedure that is sometimes referred to, in the context of AI, as 'heuristics', which should be included under the present heading. Heuristics refer to 'judgements' or 'rules of thumb' that may be included in a computer program, but that do not provide definitive mathematical solutions to the problem at hand. For example, in a chess-playing computer program, the computer may be instructed to analyse positions by calculating the results of all possible alternative successions of moves by each player, a good many moves deep. It would not be possible to take this analysis as far as checkmate (or as far as some accepted winning position) in most cases, there being far too many alternatives to work through before an adequate depth of move can be achieved for this. Instead, the various strands of computation would be terminated as soon as some appropriate 'heuristic' criterion is satisfied. The resulting program would not play perfect chess, but nevertheless it could be made to play chess extremely well, and some such programs – most particularly 'Deep Thought' constructed by a group at Carnegie Mellon headed by Hsiung Hsu – can defeat all but a very small minority of human players.

Computers that act according to such procedures are still acting algorithmically, of course, even though their action is not given by some algorithm that is *guaranteed* to provide the desired result (e.g. to play perfect chess). If such procedures were applied to the solving of mathematical problems as a whole, then they could be thought of as giving an algorithm such as Y, and it might be imagined that this is indeed how mathematicians actually operate. However, a *known* procedure of this type would certainly not be accepted by the mathematical community as providing an adequate criterion of mathematical proof. Mathematicians require a degree of rigour that makes such heuristic arguments unacceptable – so no such known procedure of this kind can be the way that mathematicians actually operate. We are left with the possibility that Y might be a horrendously complicated unknowable algorithm, and we are really back with the situation that we had before, with the unknowable algorithm X (but with Y as actually an inferior inaccurate version of X), and the same objections will apply as before. Most particularly, the arguments from natural selection are at least as damning as they were before.

3 Do we use an ever-changing algorithm?

There is, however, the possibility that we might be using an algorithm, perhaps like X or Y, but that is continually changing with time by being improved with experience. But now we must ask, as we did briefly before: by what procedures is our putative algorithm supposed to improve itself. If these very procedures are things that could be programmed on a computer, then they are themselves algorithmic. All the arguments that we have been considering before should really be applied to the *entire* algorithm controlling the computer's action, including that part which governs its mode of change. This algorithm includes all the 'learning procedures' that have been laid down at the beginning. If the only ways in which the algorithm can change is by such pre-assigned algorithmic means, then we do not, strictly speaking, have a changing algorithm at all but just a single algorithm, say X or Y, as before and all the objections raised above will still apply.

There is a particular type of learning system that has gained considerable popularity in recent years, referred to as a ***neural network***. This type of system is based initially on certain ideas about how brains actually work – most particularly, the idea of ***brain plasticity***, according to which the strengths of the various connections between neurons in the brain are continually being modified in order to improve performance – but in a 'neural network' the system is implemented entirely electronically. At the present time, such systems are normally run on ordinary general-purpose electronic computers, but the intention is that, eventually, special-purpose electronic hardware would be constructed, so that a much greater speed and efficiency can be achieved. (It should be mentioned that the biological mechanisms whereby the strengths of connections between neurons in the brain are *actually* modified are largely unknown. Neural networks use their own procedures, and these probably differ in important ways from what happens in the brain plasticity of actual brains.) Neural networks have been applied to many varieties of different task, such as the recognizing of faces, or distinguishing the quality of different sounds, or prospecting for mineral deposits, or controlling a robot arm. They can be effective for tasks where understanding seems not to be playing a crucial role – so it is hard to see that they could be of much use in forming mathematical judgements.

A basic feature of a neural network, distinguishing it from the normal type of algorithmic programming, is that rather than having a previously given algorithm that is specifically provided in order to solve some particular problem or class of problems (which would sometimes be referred to as the normal 'top-down' organization) we are, instead, initially provided with merely a loosely connected family of units (the electronic 'neurons') where the strengths of the various connections are continually being modified in order to maximize the quality of the output. In this way the system continually 'learns', so that the output improves all the time. The action of such a system is still algorithmic, since it can be implemented on an ordinary general-purpose electronic computer, but there is a crucial difference in the intentionality underlying the algorithm. Now, the algorithm according to which the system acts at any one time is not one that has been fed in initially in order to provide a specific solution to a pre-assigned problem, but it is one that has gradually evolved on the basis of improving the quality of its output ('bottom-up' organization).

Nevertheless, the procedure taken as a whole still constitutes an algorithm, provided that the judgement as to the quality of the output at any given stage is made according to some (possibly 'heuristic') algorithmic criterion. This would indeed normally be the case if one is thinking of using the neural network to solve some clearly specified problem. Certainly, if one attempted to use a neural network to solve problems in pure mathematics, then its action would be entirely algorithmic. Thus, we would again have a situation of the type already considered under the previous heading: an approximate algorithm Y; and the various objections to this being *actually* the procedure that mathematicians use would apply as before.

4 *Is the role of the environment essential?*

Of course, it might be the case that it is the continual input that we get from the environment that gives our mathematical understanding its (apparently?) non-algorithmic character. It is undoubtedly the case that mathematicians are indeed being continuously influenced by their environment. Even the very idea of what we mean by a natural number could hardly have formed in our minds were it not for our experiences with various different manifestations of such numbers – at least of the smaller ones: 'three oranges', 'two socks', 'ten pennies', etc. Moreover, although advanced mathematics need not have any very precise correspondence with our immediate experiences, and mathematical reasoning is wholly abstract, mathematicians frequently do gain their inspirations from everyday things in the world around them.

But is this input from the environment something that has an essentially non-algorithmic character? To put the question another way, might it be that it is in principle impossible to simulate the environment computationally? The question is not whether or not the actual environment of a particular individual could be precisely simulated, but whether or not everything that might be relevant to the building up of an individual's mathematical understanding could be computationally simulated. I find it hard to see where there can be any essentially non-algorithmic input coming in here. In any case, if one takes refuge in the possibility that it is the environment that supplies the needed non-algorithmic ingredient in our mathematical thinking, then one is left with the problem of explaining how the physical environment can behave non-algorithmically. If one is prepared to believe that the physically governed external world could behave in a way that cannot be simulated computationally, then one must surely also accept that a (physically controlled) human brain might behave non-computationally. This after all, is what the thrust of my argument is intended to show.

5 *Do random ingredients change the discussion?*

In the discussions so far I have been considering 'algorithmic action' to be essentially that which could be carried out in principle by a modern general-purpose electronic computer (with an unlimited store) – which, technically, means by a Turing machine. The action of a Turing machine is always completely determined in advance, even though the ultimate behaviour of the machine might be hard to predict. One might expect that some essential differences could arise when we relax this condition, and allow the presence of random ingredients to be introduced into the calculations. In modelling the environment, in particular, one would normally expect that there would be many uncontrollable parameters, and these would have to be modelled in some way in terms of a random input. Moreover, in the action of the brain itself, it is to be expected that there would be a great deal of random activity, and one could imagine that this could have important implications with regard to its overall performance.

However, in computational simulations of physical systems, the presence of random ingredients poses no problem, provided that one is not asking for a prediction of the actual detailed way in which a particular physical system is going to behave, but is merely asking for a typical example of the way in which such a physical system *might* behave. The latter, after all, is all that we would be asking of an artificially intelligent system. There are many algorithmic means of producing an *effectively* random input. In place of a 'genuinely' random sequence of digits, the computer produces a sequence of what are called 'pseudo-random' digits which, for all intents and purposes, behaves as a random sequence would, despite the fact that the sequence is generated entirely algorithmically. Thus, in practice, it does not seem to be any significant restriction to limit one's attention to entirely algorithmic computation.

In present-day physics, there are two distinct ways in which (apparently) random behaviour can result in a seemingly well-specified physical system. In the first place, there is the phenomenon known as *chaos*, whereby a completely deterministic classical system can, in practice, behave as though it were not deterministic at all. This is because the accuracy according to which the initial state needs to be known, for a deterministic prediction of its future behaviour, can be totally beyond anything that is conceivably measurable. An example that is often quoted in this connection is the detailed long-range prediction of the weather. The laws governing the motion of air molecules, and also the other physical quantities that might be relevant, are all perfectly well known; however the weather patterns that may actually emerge, after only a few days, depend so subtly on the precise initial conditions that there is no possibility of knowing these conditions accurately enough for reliable prediction. Of course the number of parameters that would have to enter into such a calculation would be enormous, so it is perhaps not surprising that prediction, in this case, might prove to be virtually impossible in practice. However, such so-called chaotic behaviour can occur also with very simple systems consisting of only a small number of particles. Imagine, for example, that one is to pocket the fifth snooker ball E in a chain A, B, C, D, E by hitting A with the cue so that A hits B, causing B to hit C, then C to hit D, and finally D to hit E into the pocket. So long as no two successive balls are initially close up against one another, the accuracy needed for this is far in excess of the abilities of any expert snooker player. If there were ten in the chain, then even if the balls were perfectly elastic perfect spheres, the task of potting the final ball would be far beyond the most accurate machinery of modern technology. In effect, the behaviour of the later balls in the chain would be random, despite the fact that the Newtonian laws governing the behaviour of the balls is mathematically completely deterministic.

From time to time, it has been suggested that this phenomenon of chaos, if it occurs in the internal action of a physical brain, might be what enables our brains to behave in ways that appear to differ from the computably deterministic activity of a Turing machine. However, if all that we can get from chaos is randomness, then this will be no use to us. For randomness itself does not give us anything usefully beyond ordinary algorithmic computation. Does chaos give us anything that is usefully non-computable, rather than just randomness? In fact the behaviour of chaotic systems can sometimes be highly structured and much more interesting-looking than purely random behaviour. It remains to be seen, however, whether this kind of structure contains anything approximating genuinely non-algorithmic behaviour. It would be very interesting to see if this is possible, but so far no one appears to have argued any kind of case that it is.

The other place where modern physics asserts that randomness arises in the precise behaviour of a physical system is in *quantum mechanics*. Many people are familiar with the fact that quantum theory does not give us a deterministic picture of the time-evolution of the world and that it merely provides probability values for the various alternative outcomes of experiments that may be performed on a system. What is not so commonly appreciated is that this probabilistic – i.e. random – behaviour does *not* occur at the submicroscopic *quantum level* of molecules, atoms and subatomic particles. For at that level, things are described as evolving according to the precise deterministic Schrödinger equation, and the behaviour is just as deterministic as it is at the *classical level* of everyday experiences, where the laws of Newton and Maxwell (and also Einstein) hold sway. Instead, the quantum randomness is effectively taken to occur at some mysterious middle-ground level, intermediate between the quantum and classical levels. A feature of the quantum level is that all the various alternative things that a system *might* do have also to be allowed to happen *together* in some mysterious kind of quantum superposition. But when the effects of some of these superposed alternatives become magnified to the classical level, in accordance with a *measurement* being performed on the system, one alternative or the other becomes realized as actuality. At this stage, Nature's choice of alternative is governed merely by probability. According to theory, what one must do – and this is all that quantum mechanics tells us about what happens during measurements – is to adopt the strange procedure referred to as *state-vector reduction* (or *wavefunction collapse*). According to this, the quantum description (the state-vector, or wavefunction) undergoes a discontinuous random change, where what happens is restricted only by the probability laws that the theory tells us how to calculate. It is here, and only here, that the deterministic character of the underlying theory is changed to a probabilistic one. At both the quantum level (Schrödinger's equation) and the classical level (Newton-Maxwell laws) the behaviour is basically deterministic (considerations of chaos notwithstanding), and the randomness comes in only when one tries to bridge the gap between the two. In this connection it should be pointed out that the difference between the quantum and classical levels is not really one of size. (Quantum effects have been observed to take place over a separation of many metres.) It would be somewhat more accurate to say that the quantum level refers to phenomena for which the differences in energy between the states under consideration are small. In fact the whole issue of what *actually* happens when the alternative effects of differing quantum possibilities are magnified to the classical level whenever such a 'measurement' takes place is fraught with controversy. But whatever one believes quantum theory is telling us about the actuality of the world, some procedure akin to the one just described must be adopted in order that the necessary probability calculations can be made.

What might be the relevance of this to the activity of a human brain? As was the case with chaos, all that we appear to obtain, when the procedures of quantum theory are adjoined to the mechanistic determinism of classical theory, is a random element that would seem to be of little use to us if we are expecting to find something usefully non-computable coming out of it all. (In fact, there are certain profoundly intriguing features of the behaviour of quantum systems that differ significantly from what can be achieved by classical systems – such as non-locality[7], and the possibility that certain calculations can, in principle, be performed more rapidly – but these do not change the class of problems that can or cannot be solved from what they were for a Turing machine.) As our understanding rests at the moment, the incorporation of conventional quantum mechanics into our discussion does not significantly help us in our search for a non-algorithmic ingredient that could be relevant to our thought processes.

Thus, as things stand, chaos and quantum mechanics both leave us with only randomness and not with something usefully non-algorithmic. There remains a hope, in each case, that future developments in theory might improve the situation. With chaos, this hope rests on the possibility that something more useful than mere randomness might arise in suitable circumstances; with quantum mechanics, it rests on the finding of a new physical theory with a more satisfactory way of handling the transition between the quantum and the classical level. Both hopes represent highly unconventional viewpoints at the present time. I shall return briefly to these possibilities under the final heading.

6 *Must we resort to mysticism?*

Since we seem, so far, to have been led into an impasse by following the methods of conventional scientific theory, is it not perhaps reasonable that we should abandon the physicalist view of the mind altogether and turn, instead, to a viewpoint more in line with some form of mysticism? Even Kurt Gödel appears to have found himself driven to a rather mystical standpoint. In commenting on Alan Turing's computational view of the mind[8], Gödel suggested that the mind need not be limited by the brain's finiteness. He seemed to have believed, along with Turing, that the objects of the physical world should behave 'mechanically' – by which he meant, in effect,'algorithmically' – but unlike Turing, he appeared to believe that the human mind was something beyond the physical brain. In this way, the mind would be able to perceive the Platonic truth of mathematical statements in some way not totally dependent on the physical action of the brain.

As I have stated in my introductory remarks, I intend to stay within the framework of scientific possibilities in addressing the issues of mind. The evidence that mental attributes result from activity in the physical brain seems to me to be overwhelming. While it is certainly true that we do not understand how mental phenomena can arise from the activity of a physical object, the case is surely powerful that somehow they can, at least when the 'physical object' is a living human brain. To my own way of thinking, the resolution of the profoundly puzzling issues raised by the evident fact that certain types of physical object can indeed give rise to conscious minds must ultimately lie in our finding a deeper understanding of what physical objects actually are. As science has progressed, our descriptions have become more and more accurate and complete; but at the same time, they have become more and more abstract and mathematical. Our very picture of concrete reality has changed to something almost totally abstract and, in many ways, mysterious. It seems that we need to turn to the abstract Platonic world of mathematical forms if we are to find where ultimately concrete reality is to find its home. The mystery of mind can ultimately be understood only in terms of the mystery of matter itself, and in the undoubted mystery of the mathematics that so precisely seems to control its every action. This is not mysticism; it is the ultimate purpose of science itself.

7 *Are there non-algorithmic physical laws?*

If we are to stay within the laws of physics in our search for a non-algorithmic action that might underlie our conscious thought processes (such as those that give rise to the mental quality of mathematical understanding), then we must find something usefully non-algorithmic in those laws. As I asserted earlier, mere randomness will not do. We need something a good deal more sophisticated than randomness if we are ever to find a possible physical basis for our conscious understanding.

What kind of physical laws could exhibit non-algorithmic behaviour? It is certainly possible to provide various different kinds of 'toy model universe' that are non-algorithmic. Such a model can even be completely deterministic, as the following (admittedly totally unrealistic) example illustrates: Let the state of our 'universe' be given by a pair of natural numbers n, q, where the (discrete) time is also given by a natural number t. The rule for the evolution of this universe is that, if the state at time t is (n,q), the state at time $t+1$ is to be $(n+1,q)$ if the calculation $C_n(n)$ stops and it is $(q+1,n)$ if $C_n(n)$ doesn't stop. Since there is no general algorithm for deciding whether a calculation $C_n(n)$ stops or not, the future behaviour of our model universe is not computable, even though it is completely deterministic. Although this particular model is not to be taken seriously as a picture of reality, of course, it is not at all impossible that the *actual* physical laws could exhibit a non-algorithmic behaviour of this general kind. There are many classes of mathematical problem that are non-computable and that could take the place of the non-stopping of $C_n(n)$ in the example just cited (the solubility of families of polynomial equations in terms of natural numbers, the topological equivalence problem for four-dimensional manifolds, the problem of deciding which finite sets of polygonal tiles will cover the Euclidean plane without overlaps or gaps, etc.). There could also be other kinds of example of a much more subtle, and possibly more realistic, nature.

There is, as yet, no indication that the phenomenon of chaos can exhibit anything usefully non-random that approximates a non-algorithmic action of this general kind. In this connection it should be pointed out that chaotic behaviour can certainly be simulated computationally, and I think that the arguments that I have given earlier argue strongly against such a computational model of conscious understanding. For reasons such as these, I think that it is unlikely that chaos can supply our needed answer.

On the other hand, there are numerous reasons, coming from physical theory itself, for believing that present-day quantum theory must someday be profoundly modified, the procedure of state-vector reduction being distinctly unsatisfactory physically. It is my own opinion that such a necessary modification will someday be discovered, and that it is here that we shall find our needed non-algorithmic behaviour – to replace or amend the probabilistic procedures that present-day quantum theory seems to require. Such a theory would need to have an even deeper, more profoundly subtle, and universal nature than the theories that science has provided to date, and its unity with the Platonic world of mathematics should also be more secure. Perhaps, then, we shall be better able to find a basis for mental phenomena within scientific explanation than has been remotely possible hitherto.

Notes

Opening lecture, Wolfson Lecture series on Artificial Intelligence; 15 January, 1991; Wolfson College, Oxford. This account is partially based on a draft of part of the opening chapter of a new book by the author, to be published, in due course, by the Oxford University Press.

1 These four alternatives are explicitly described in, for example, Johnson-Laird (1987), p. 252. (It should be noted, however, that what he refers to as the 'Church-Turing thesis' is not given altogether appropriately, and its relevance to 'viewpoint 3' requires further

discussion.) For a spectrum of opinions on the mind-brain-computer issue, see also various other articles in Blakemore and Greenfield (1987).

2 Johnson-Laird (1987) attributes this viewpoint originally to Craik (1943). Other proponents appear to include Alan Turing(?), Allen Newell, Herbert Simon, John McCarthy, Marvin Minsky, Douglas Hofstadter and Hans Moravec.

3 See the discussion in Moravec (1988), chapter 6.

4 From his explicit writings, I take this to be the originally stated viewpoint of Searle (1980).

5 See, for example, Searle (1980).

6 For earlier arguments in support of the thesis that Gödel's theorem shows that (mathematical) thinking cannot be computational, see Nagel and Newman (1958), Lucas (1961). There have been numerous counter-arguments to these; see, for example, Benacerraf (1967), Lewis (1969, 1989), Hofstadter (1979).

7 See Bell (1987).

8 See his collected works: Gödel (1990).

References

Bell, J. S. 1987: *Speakable and Unspeakable in Quantum Mechanics.* Cambridge, Cambridge University Press.

Benacerraf, P. 1967: God, the Devil and Gödel. *The Monist*, **51**, 9–32.

Blakemore, C. and Greenfield, S., eds 1987: *Mindwaves: Thoughts on Intelligence, Identity and Consciousness.* Oxford: Basil Blackwell.

Craik, K. 1943: *The Nature of Explanation.* Cambridge: Cambridge University Press.

Gödel, K. 1990: *Kurt Gödel, Collected Works, Vol. II (publications 1938–1974).* (Edited by S. Feferman, J. W. Dawson, Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, J. van Heijenoort) New York: Oxford University Press.

Hofstadter, D. R. 1979: *Gödel, Escher, Bach: an Eternal Golden Braid.* Hassocks, Sussex: Harvester Press.

Johnson-Laird, P. 1987: How could consciousness arise from the computations of the brain? In Blakemore, C. and Greenfield, S. (eds), *Mindwaves: Thoughts on Intelligence, Identity and Consciousness.* Oxford: Basil Blackwell.

Lewis, D. 1969: Lucas against mechanism. *Philosophy*, **44**, 231–3.

Lewis, D. 1989: Lucas against mechanism II. *Can. J. Philos.*, **9**, 373–6.

Lucas, J. R. 1961: Minds, Machines and Gödel. *Philosophy*, **36**, 120–4;

reprinted in Alan Ross Anderson 1964: *Minds and Machines*, Englewoods Cliffs: Prentice Hall.

Moravec, H. 1988: *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, Mass./London: Harvard University Press.

Nagel, E. and Newman, J. R. 1958: *Gödel's Proof.* London: Routledge & Kegan Paul Ltd.

Penrose, R. 1989: *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press; see also summary, commentaries and author's response in *Behavioral and Brain Sci.*, **13** (1990) 643–705.

Searle, J. R. 1980: Minds, Brains and Programs. In *Behavioral and Brain Sci.*, **3**, reprinted: *The Mind's I*, D. R. Hofstadter and D. C. Dennett (eds) 1981: Hassocks, Sussex: Harvester Press.

217

THE MODERN PHYSICIST'S VIEW OF NATURE

The concept of nature: The Herbert Spencer Lectures. Ed. J. Torrance. Clarendon Press, Oxford, 1992

The modern physicist's view of nature

ROGER PENROSE
Mathematical Institute, University of Oxford

I have been asked to describe the modern physicist's view of nature, and you may very reasonably ask: what right does a mere mathematician have to pontificate on what he believes to be the picture of nature held by physicists? Perhaps it would be best if you are to take my remarks as presenting the *mathematician's* view of the modern physicist's view of nature, or rather, one particular mathematician's view of the modern physicist's view of nature. None the less, it is certainly true that modern physics could not exist without mathematics, nor, indeed, without mathematics of ever-increasing sophistication. This has been particularly true of work done over the past ten years or so, but it has been true, also over the entire history of the development of successful physical theory from early Greek times until the present day. Perhaps, after all, it is not so unreasonable for a mathematician to take over the task of describing the physicist's view of nature.

Let me begin by attempting to put the physicist's achievements in some kind of perspective. First of all, I think that it is important for us not to get too carried away by their excitement about whatever fashionable ideas happen to be taking their interest at any one moment. For the physicists themselves, on the other hand, it is not at all inappropriate that they should get carried away. Their own enthusiasm is a vital ingredient for real progress in fundamental physics. Physics is an area where the mathematics can get extremely difficult at times and is frequently very tedious, so that a physicist's feeling that goals (and perhaps rich rewards) are close at hand is essential for progress. For the rest of us (and for the physicists, also, when in their more reflective moods), it is necessary to be more circumspect. Let us try to assess the status of the physicists' achievements over the whole of history.

It should be pointed out that some of these achievements have indeed been truly remarkable. The best of the physicist's theories have acquired a status unmatched in any other science (where by the term 'science', I do not include mathematics). Such theories I refer to as SUPERB. To qualify as SUPERB, it is not necessary for the theory to apply without refutation to the phenomena of the world, but its range and accuracy must, in some appropriate sense, be *phenomenal*. The way that I am using the term 'SUPERB', it is a quite extraordinary fact that there are any theories in this category at all. I do not believe that any basic theory in any other science could properly qualify as SUPERB. The only possible exception might be Darwin and Wallace's theory of natural selection (as amended by Mendel), but even that theory, for all its power and scope, and for all the insights that it gives us into the workings of the biological world, perhaps does not quite achieve the precision in numerical accuracy that would allow it to qualify, in the terms that I am laying down, for the SUPERB category. In saying this, I am in no way being disrespectful to the theory of natural selection. On the contrary, I am simply emphasizing the extraordinary accuracy and scope of those physical theories that to indeed qualify as SUPERB.

Which are the SUPERB theories? Let me list these as I see them:

(1) Euclidean geometry
(2) classical statics
(3) Newtonian dynamics
(4) Maxwellian electrodynamics
(5) special relativity
(6) general relativity
(7) thermodynamics
(8) quantum mechanics
(9) quantum electrodynamics.

I shall give a brief description of each of these, indicating why it should indeed be considered as a SUPERB physical theory, and then touch upon the rest of mainstream physics, all of which belongs to the remaining categories, which I label USEFUL or TENTATIVE (or perhaps even MISGUIDED).

EUCLIDEAN GEOMETRY

Let us begin with Euclidean geometry. This is simply the subject of 'geometry' that we learn something of at school. Why do I refer to Euclidean geometry as a *physical* theory rather than a branch of mathematics? Of course it is mathematics as well: the classical example of a deductive system, where axioms are laid down to begin with and hosts of propositions are then derived from them in step-by-step logical ways. But the reason for the very choice of axioms in the first place is that they agree extraordinarily closely with the physical notion of a 'point' as the tiniest of particles and of a 'straight line' as a ray of light—or else as a line attached to a rigid body (say a ruler), where the line can be moved along itself and independently turned about itself. If we had had no familiarity with such physical straight lines, with rigidity, or with tiny point-like specks of physical matter, then Euclid's geometry would not seem so 'natural' or 'obvious' to us. The seeming self-evidence of the Euclidean axioms arises from the fact that we inhabit a physical space occupied by rigid physical objects that behave in a very accurate way according to the laws of Euclid's geometry. It is initially the structure of the physical world, not just mathematics, that suggests to us the axioms that we should choose; it is mathematics, on the other hand, that is concerned with deductions that can be made from those axioms.

There are two distinct ways in which the ideal Euclidean theory of geometry can fail to describe the physical world precisely. The first of these is that physical bodies need not be precisely rigid; nor can we expect to find, in nature, precisely straight physical lines of zero thickness or tiny specks or marks that are precisely pointlike. Normally one does not regard this as an essential obstacle to our taking the geometry of 'actual' space to *be* Euclidan. We may regard physical objects as having finite extension, and also as being deformable, so long as these extensions and deformations are things that take place *in* Euclidean space and can be measured and unambiguously described, according to an appropriate theory, in terms of standard Euclidean concepts of distance and angle—in short, so long as we can, in principle, make the appropriate *corrections* that tell us how we must expect our actual physical objects to deviate from the physical ideals of 'pointlike', 'straight', or 'rigid'. Accordingly, we think of 'space' itself as having a precisely Euclidean structure, the physical objects themselves perhaps falling short of this ideal in one way or another.

The second way in which Euclidean geometry might fall short of complete accuracy is more subtle. There are actually kinds of geometry other than Euclidean geometry that are mathematically possible. One of these, for example, is what is called *Lobachevskian* (or hyperbolic) geometry. Lobachevskian geometry satisfies all the axioms of Euclidean geometry except for 'Euclid's fifth postulate'—with the implication that in Lobachevskian geometry there will be more than one straight line through a given point that is parallel to another given straight line (not through the point). In Fig. 1, I have provided an Escher print giving a representation of this geometry (within a portion of a Euclidean plane). All the black fish are to be regarded as congruent and so also are all the white fish. The outer circle denotes 'infinity', and the entire Lobachevskian plane is represented as the interior of this circle. The Lobachevskian straight lines are represented as circles meeting the boundary circle at right angles. We can also think of Lobachevskian *three*-space as being represented in a similar way, except that the space is now represented as the interior of a sphere instead of a circle. (These representations were found by Poincaré—although I am told that Beltrami had known about them many years earlier.) Now it could turn out that the most appropriate concept of 'actual physical space' is not Euclidean at all, but it might be some other kind of geometry, such as the Lobachevskian geometry just described. We might still try to imagine that the underlying space is 'actually' Euclidean, and that any deviations from Euclidean geometry that *real* physical bodies might experience could be accounted for by taking the 'corrections' point of

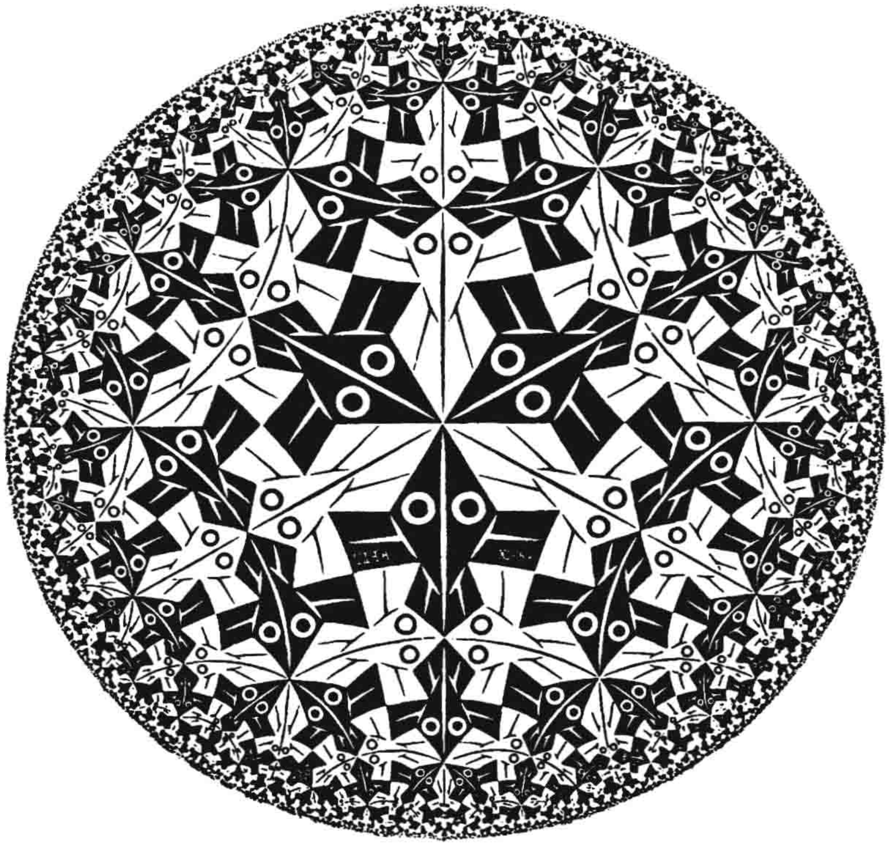


Fig. 1 Escher's depiction of Lobachevskian space. (Think of all the black fish as congruent and all the white fish as congruent.)

view, as described above. However, this may not be at all appropriate. In the first place, deviations from the supposed 'actual Euclidean underlying space' that real physical bodies would need to acquire can get unreasonably large, perhaps getting larger and larger and farther into the distance that these bodies are examined, so that the supposed underlying geometry begins to have little relationship with the (approximate) geometry described by the real bodies themselves. A second reason for the inappropriateness of such a concept of 'underlying Euclidean geometry' is that there may be no uniqueness about it when it does not accord closely with the geometry respected by real physical objects. There may be a clear sense in which the most appropriate physical geometry is not Euclidean after all.

How well does Euclid's geometry in fact describe the physical world? It turns out that when we have taken into account all the 'corrections' that need to be made, there still remains a deviation in the behaviour of physical objects, from the underlying Euclidean ideal. This deviation is utterly tiny, however, and in total it would not amount to more than the diameter of an atom of hydrogen in a region a metre or so across in any normal part of space. But the fact that deviations do occur emphasizes that Euclidean geometry is indeed a falsifiable physical theory and not just a piece of mathematics. The extraordinary accuracy of this theory, and the scope and power of its applicability to the physical world gives it an undoubted place among the SUPERB theories. The fact that deviations do exist in no way bars it from membership. (We shall see with the other SUPERB theories, that the existence of tiny deviations is the rule rather than the exception.)

The deviations from Euclid's geometry are in fact accounted for by anther SUPERB theory, general relativity, which we shall be coming to shortly. But before we move on, I should point out that there is a fundamental ingredient of Euclidean geometry, namely the concept of a *real number*, that was introduced, in effect by Eudoxos in the fourth century BC, that has remained with us, unchanged to the present day, as an essential ingredient of *all* successful physical theories including general relativity. (Real numbers are, in modern everyday terms, what we would describe using infinite decimal expansions—although Eudoxos, in his day, used a different method.) Whether the real number will retain this essential status in future SUPERB theories remains to be seen, however.

CLASSICAL STATICS

There was another great scientific achievement established largely by the ancient Greeks, although a little later than Euclidean geometry. This was the theory of *statics*, due mainly to Archimedes (287–212 BC) and Pappos (late third century AD), and also

Simon Stevin (*c*. 1548–*c*. 1620). This theory described objects in equilibrium (i.e. balanced, so that there is no motion), and it included the law of the lever and properties of bodies suspended in fluids, including the concepts of pressure and force, etc. The accuracy of this theory must have been very great, especially considering the antiquity of its basic ideas, and bearing in mind the comparatively primitive means that were available to check them. The ancients could construct such an accurate body of laws largely because they were able to find a number of ingenious and clearcut mathematical arguments, starting from certain 'obvious' principles (such as symmetry properties, and the non-existence of a *perpetuum mobile*). We do not now tend to think of classical statics as a physical theory in its own right, because it has been subsumed by the more general scheme of Newtonian dynamics. But I have chosen to list it separately here because, for some eighteen centuries, it stood independently as a profound testament to the power of mathematical theory in accurately describing the physical world.

NEWTONIAN DYNAMICS

The main thing that was lacking in early Greek science was a good understanding of the laws governing bodies in *motion*. They appear to have believed that the laws governing the motions of celestial bodies were quite different from those which controlled the motions of bodies on earth. Their ideas about earth-bound motions were very inaccurate. However, they did have accurate empirical knowledge of the motions of celestial bodies, and they formulated a family of laws to describe such motion, now referred to as the *Ptolemaic* system (due, in successive refinements, to Eudoxos, Hipparchos, and Ptolemy). This system described the movement, around the earth, of the planets, the moon, and the sun, in terms of complicated compositions of circular motions. It provided a useful scheme of quite creditable accuracy and predictive power, though it became more and more over-complicated as the need for greater accuracy increased. It was a wonderful classical example of a theory belonging to the category that I am referring to as USEFUL (for about twenty centuries, in fact!). The work of Copernicus and Kepler changed that picture considerably, and it provided another USEFUL theory, according to which the planets move, instead, about the sun, in elliptical orbits with the sun at one focus.

The system of dynamics initiated by Galileo in about 1600 and completed by Newton in 1687 changed things dramatically. Now, the same laws were seen to apply here on earth as governed the motions of celestial bodies. Moreover, instead of having a merely descriptive scheme in which the orbits described by the planets' motions were thought of as geometrical things in their entirety, we were presented with a full *dynamics*, where laws were given determining how the motion of each body changed from moment to moment—i.e. how each body *accelerated*—in terms of the locations and nature of all other bodies. Newton's dynamical laws, together with specific laws of force such as the inverse square law of gravitation, gave an incredibly detailed and accurate account of the motions of bodies. Now it was possible not only to *deduce* Kepler's elliptical motion from more primitive dynamical laws, but innumerable corrections to this motion could be worked out and found to be in very close accordance with what was actually observed. The agreement, in Newton's day, between theory and observation was to something like one part in a thousand; today, that agreement is to better than one part in ten million. Certainly Newtonian dynamics must qualify as SUPERB!

It should be pointed out that the much earlier SUPERB theory of Euclidean geometry is an essential part of Newtonian dynamics, this geometry forming the background to all Newtonian descriptions. Moreover, not only has the SUPERB theory of classical statics been subsumed into Newtonian dynamics, but so also has Kepler's USEFUL theory of planetary motion; while, on the other hand, that other USEFUL theory, the Ptolemaic system, has suffered the opposite fate: it has faded away altogether.

MAXWELLIAN ELECTRODYNAMICS

Classical Newtonian theory is concerned with the motions of *particles*, and a single particle is something that can be described by a small finite number of parameters (three to determine its position in space and three more to fix its velocity, or momentum; and one may require a few more for its mass and, say electric charge and perhaps radius and spin). Classical particles are not sufficient for the description of nature, however. Since the time of Faraday (1791–1867) and Maxwell (1831–79) it has become clear that electric and magnetic fields are things that must be considered as having physical existence in their own right, and cannot be regarded as merely appendages to the physically existing particles, as had been possible with the 'gravitational field' of Newtonian theory, which can be regarded as merely providing 'bookkeeping' for the various forces on the particles. The Faraday–Maxwell theory explains the phenomenon of *light* as a propagation through space of disembodied electric and magnetic fields. Such electromagnetic waves can carry energy and information from one place to another at the speed c that light indeed achieves. Not only is visible light now understood in these terms, but a host of other phenomena are explained and predicted on the basis of this theory, such as radio waves and X-rays. Electromagnetic fields must be regarded as just as 'real' as particles themselves.

Maxwell provided a beautiful family of equations to describe the behaviour of electromagnetic fields. Maxwell's equations are examples of what are called *partial* differential equations, which distinguishes them from the *ordinary* differential equations that arise in the Newtonian descriptions for the motions of individual particles. This difference arises largely because Maxwell's field requires a (continuous) *infinite* number of parameters in order to describe it completely, as opposed to the finite number that is needed for Newton's particles.

Maxwell's theory is accurately valid over an extraordinary range, reaching inwards to the tiny scale of atoms and subatomic particles, and outwards, also, to that of galaxies, some 10^{36} times larger. (At the very small end of the scale, Maxwell's equations must be amended appropriately by combining them with the rules of quantum mechanics.) Maxwellian dynamics, also, must qualify as SUPERB.

SPECIAL RELATIVITY

One of the most important properties of the dynamical theory of Galileo and Newton is that it satisfies a *relativity* principle. The dynamical laws remain unchanged if the reference frame to which they are referred is replaced by another frame in uniform (i.e. non-accelerating, non-rotating) motion with respect to the first. Thus, no local experiment can distinguish a state of uniform motion from the state of rest. (This property is essential, in order that the Copernican heliocentric viewpoint can make dynamical sense.)

A fundamental difficulty arises, however, when Maxwell's electromagnetic theory is adjoined to Newtonian dynamics. Maxwell's equations do not share the invariance properties relevant to the relativity principle of Galileo and Newton. (The simple reason for this is that the speed of light c is built into Maxwell's equations, so that light propagation has an unambiguous velocity in the original reference frame. With respect to a moving reference frame, the speed of light would be diminished in one direction and increased in the opposite direction.) Maxwell's theory on its own, however, does satisfy a relativity principle, though the transformation laws are different from those of Newtonian dynamics. Lorentz, Poincaré, and Einstein showed how Newtonian theory must be modified in order that this new relativity principle can apply to physical systems generally. The theory thus obtained is called *special relativity*.

The most fundamental difference between the laws of special relativity and those of Newtonian dynamics is in the way in which *time* is to be understood. In Newtonian theory, there is a universal time, describable in terms of a single real parameter t. It makes invariant sense to say, of two spatially separated events, whether or not they occur at the same time. In special relativity, this is no longer true. Two events P and Q may be judged to be simultaneous with respect to one reference frame, while with respect to a second one, moving uniformly with respect to the first, one of P or Q may be judged as occurring earlier than the other.

In order to understand this curious behaviour, it is best to take a *space–time* viewpoint rather than a merely spatial one. In a space–time picture, the different points are to represent different *events*, where an event is something with a precise ('instantaneous') temporal location as well as a precise spatial location. Since the picture has to include time as well as space, it has to be *four*-dimensional; but for visualization purposes, it is usual to suppress one (or two) of the spatial dimensions. First, let us see how the relativity of Galileo and Newton can be understood in space–time terms. In Fig. 2, I have drawn a picture of Galilean space–time (with one spatial dimension suppressed). The different horizontal slices through the picture are to represent the whole of (Euclidean) space at various different times. Thus, two events are *simultaneous* if they lie on the same horizontal slice. Time is depicted as increasing upwards in the picture, so that if two events are not simultaneous, then the later one is the one depicted as lying on the higher of the two horizontal planes. The principle of Galilean relativity manifests itself in the fact that we do not make a natural association between the points of one horizontal plane and the points of the next one (so we do not draw in vertical lines indicat-

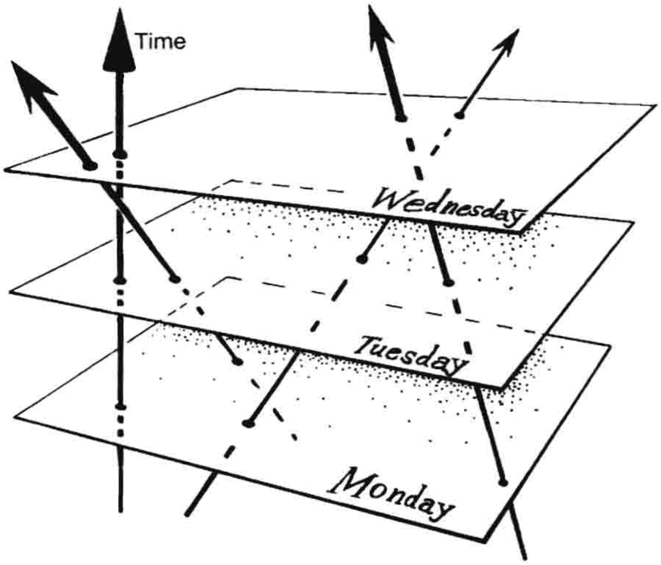
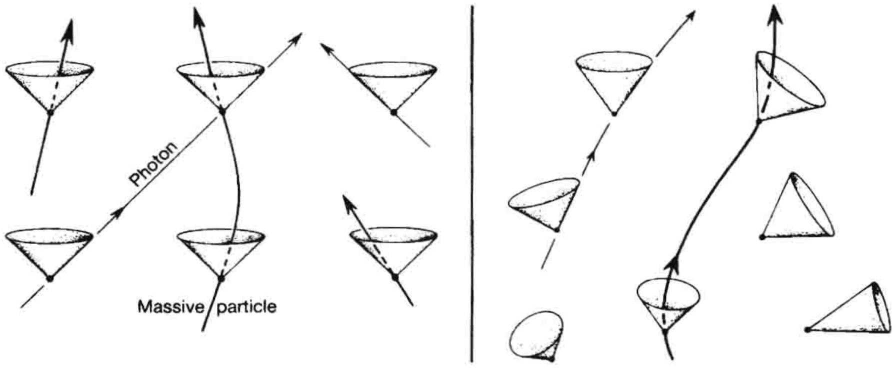


Fig. 2 Galilean space–time: particles in uniform motion are depicted as straight lines.

ing which point in space at one particular time is supposed to be identified with some particular point at another time). This illustrates the fact that there is no natural 'state of rest' distinguished from all other uniform motions. The Galilean relativity transformations (which leave the dynamical laws unchanged) correspond to 'shearing motions' of space–time, in which the horizontal planes are slid along one another, but in such a way that the straight lines in the picture (representing the histories of inertially moving particles) remain straight.

In special relativity the picture is different. We must adopt a space–time geometry that was put forward by Minkowski in 1908. There is no absolute concept of simultaneity, so we now have no horizontal planes. Instead, we must depict the *light cones*, which represent the histories of the flashes of light emitted from various events depicted in the space–time (see Fig. 3(a)). There is a concept of time that is *not* now universal, but which depends upon the motion of the object that is measuring or experiencing that time. The history, or *world-line*, of a particle—which is a curve in space–time—carries a measure of 'length', called its *Minkowskian length*, that describes the passage of time that the particle experiences. Minkowskian length is closely analogous to Euclidean length, but there is a sign difference in the defining expression,



(a) (b)

Fig. 3 Space–time pictures in relativity. The light cones at various points are depicted. (a) Minkowskian geometry; (b) general relativity.

which implies that this 'length' gets shorter and shorter as the slope of the line approaches that of the light cone (i.e. 'moving clocks run slow'), becoming actually zero *on* the light cone. Thus photons (particles of light—which we may think of as tiny localized regions of electromagnetic oscillation) experience no passage of time at all between the event of emission and the event of absorption. Any material (massive) particle must have its world-line constrained to lie *inside* the light cones. It turns out that the straight world-lines—describing inertially moving particles—are curves that *maximize* (rather than minimize) the Minkowskian length.

The transformations of special relativity (called Lorentz or Poincaré transformations) are those that preserve the Minkowskian length. They correspond to changing from the reference system of one inertially moving observer to that of another. The planes (actually Euclidean three-spaces) that appear to represent 'space' to one such observer, at each different time, can be drawn as a family of horizontal slices through the space–time picture. These are his *simultaneous* spaces. When we pass to a second inertial observer, we find that his simultaneous spaces are 'tilted' with respect to those of the first observer, so the concept of 'simultaneity' differs between the two. In particular, there could be two events R and Q, where according to one observer, R precedes Q, while for the other, Q precedes R (see Fig. 4).

Normally, we think of these differences in temporal ordering to be very tiny, or to be relevant only for relative velocities that are

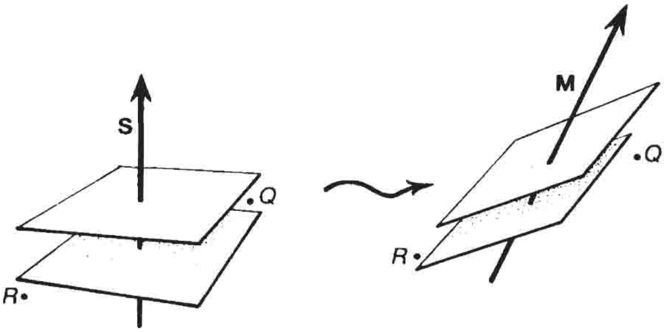


Fig. 4 The left diagram depicts simultaneous spaces for the observer **S** and the right diagram those for the observer **M**. Note that **S** thinks that *R* precedes *Q*, whereas **M** thinks that *Q* precedes *R*.

close to the speed of light. However, it is worth pointing out that even for very modest relative velocities, significant differences in time ordering will occur for events at great distances. Imagine two people walking slowly past each other in the street. The difference between two events on the Andromeda galaxy (which is the closest large galaxy to our own Milky Way, being some 2×10^{19} kilometres distant from us), that would be judged by the two people to be simultaneous with the moment that they pass one another, could amount to a difference of several days (see Fig. 5). For one of the people, the space fleet launched with the intent to wipe out all life on the planet Earth is already on its way; while for the other, the very decision as to whether to launch that fleet has not yet even been made!

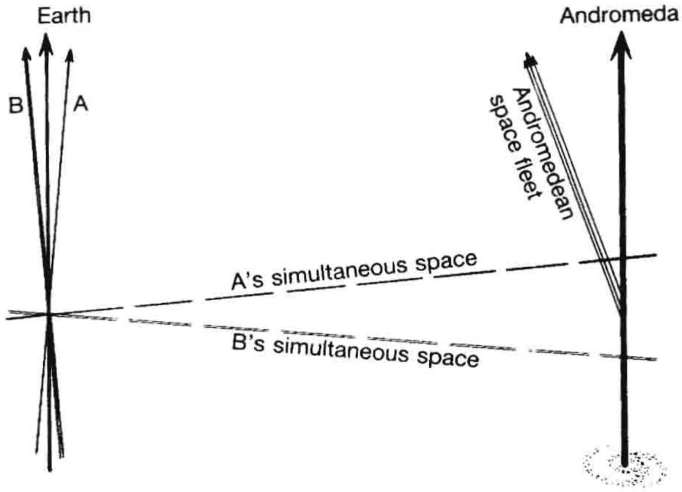


Fig. 5 Two people A and B walk slowly past each other, but they have differing views as to whether the Andromedean space fleet had been launched at the moment they pass each other.

Despite all this strangeness, special relativity is a phenomenally accurate theory. The Minkowskian length has been confirmed in many different ways as the 'time' that clocks or particles actually measure or respond to. The most direct verification of this comes from time measurements made by very accurate clocks carried in aeroplanes. These clocks 'run slow' in a way that is in precise agreement with relativity (though when the effects of gravity become important, some additional effects of *general* relativity must also be taken into account; see next section). Even more precise measurements come from observations that are being made all the time in connection with particle physics where, for example, the decay lifetimes of fast-moving cosmic rays are enormously extended with respect to a reference frame fixed on the earth. Einstein's $E = mc^2$ (which is an essential part of special relativity, arising when the Lorentz transformation laws are applied to the concepts of energy, momentum, and mass) has been confirmed in many detailed ways, to great numerical accuracy. Special relativity is undoubtedly a SUPERB theory.

GENERAL RELATIVITY

The accuracy of special relativity is not perfect, however, and tiny discrepancies arise when the phenomenon of gravity becomes important. These discrepancies are admirably accounted for by Einstein's *general relativity*. Again, we must adopt a space–time view, rather than a purely spatial one, but now the picture does not have the uniformity of Mikowski's space–time picture of special relativity, as depicted in Fig. 3. Instead, the geometry is more complicated, though it resembles Minkowskian geometry on a small scale approaching it more and more closely as the scale gets smaller. Thus, as before, we have light cones emanating from the various events in the space–time, representing the histories of light flashes originating at these events. Material particles are constrained to have their world-lines lying within these cones, whilst photons' world-lines lie along the cones. However, the cones are now not distributed in a uniform manner (see Fig. 3b), but are 'bent away' from their standard Minkowskian positions. We must think of these cones as marked on some kind of deformable medium (like a rubber sheet) so that any one of the cones can be moved, if desired, into a standard vertical position—so, locally, the space–time geometry resembles that of Minkowski. But it turns out not to be possible to distort the picture as a whole, in this way, so that

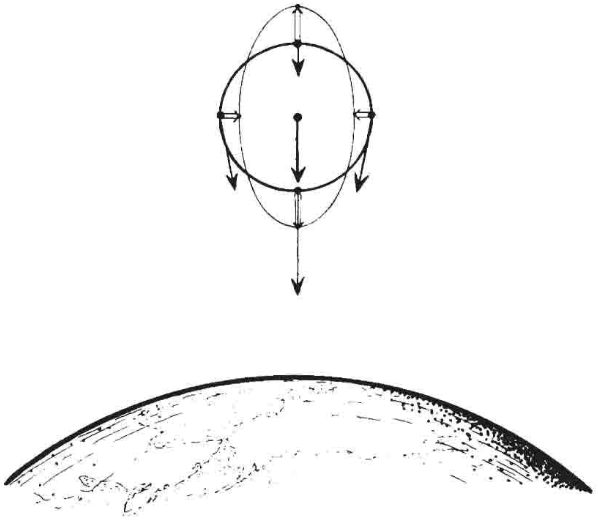


Fig. 6 The effect of Weyl space–time curvature (tidal effect). Double arrows show relative acceleration.

it exactly coincides with Minkowski space. The obstruction to our being able to do this is what is known as *space–time curvature*.

How are we to understand this 'space–time curvature'? In ordinary physical terms, space–time curvature manifests itself in the *tidal* effect of gravity. Imagine oneself situated in space, in free motion under the earth's gravitational field. Imagine, also, that one is surrounded by a sphere of particles, small compared with the size of the earth, but large enough that the slight differences in the earth's field at different points on the sphere are significant. Take the particles to be initially at rest relative to oneself (though all may share one's orbital motion about the earth). Initially, the sphere will seem to remain stationary, since everything falls together in the earth's field (in accordance with Galileo's principle that 'everything falls with the same speed in a gravitational field'); but as soon as the slight differences in the direction and strength of the earth's field on the various particles becomes apparent the sphere will begin to distort into an ellipsoid, stretched in the vertical direction (see Fig. 6). (This is precisely the same effect that the moon's gravitational field has, in raising tides on the earth—where now 'oneself' in the above descriptions is replaced by the earth and the earth by the moon.) In space–time terms, the situation is as depicted in Fig. 7. Notice that the world-lines of the surrounding particles appear to bend towards, and also away from, the central world-line (which represents 'oneself').

The situation is closely analogous to what happens with geodesics drawn on a curved surface. A geodesic is a curve of locally minimum length constrained to the surface. (Imagine a string stretched over the surface.) On a surface of positive curvature (like the surface of a sphere or ellipsoid) close pairs of geodesics that start out parallel will begin to bend inwards towards each other, whereas on a surface of negative curvature (like a saddle surface), they will begin to bend outwards away from each other (see Fig. 8). In our space–time picture of Fig. 7 we have something very similar but where effects of positive and negative curvature are both present at the same time (which can occur here because of the higher dimensionality). In fact the world-lines of particles

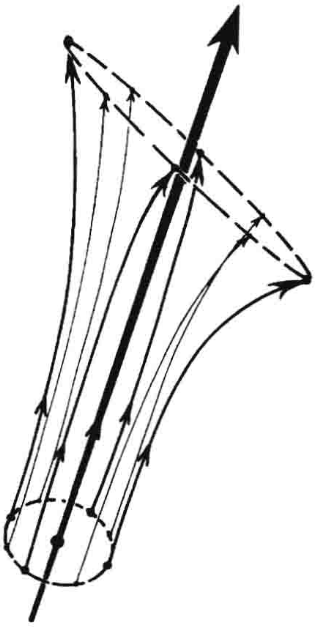


Fig. 7 Space–time curvature: the tidal effect depicted in space–time.

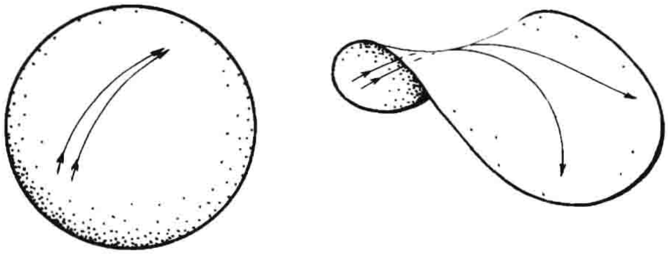


Fig. 8 Geodesics on a curved surface. With positive curvature the goedesics converge, whereas with negative curvature they diverge.

freely moving under gravity *are* geodesics in space–time, where the 'length' involved is the Minkowskian type of length that measures the *time* experienced by a particle—generalized, now, from the flat Minkowskian expression of special relativity in the same way that the length measure along the curved surfaces of Fig. 8 generalizes the length measure in a Euclidean plane. (This Minkowskian time measure is referred to as the 'space–time metric'.) Because of the sign difference in passing from a Euclidean to a Minkowskian type of length, however, it turns out that, in space–time, the world-lines of particles that are freely moving under gravity are geodesics in the sense that they locally *maximize* rather than minimize the time experienced by the particle.

The situation depicted in Fig. 6, which occurs in a region of vacuum, has a very special property: the *volume* of the ellipsoid into which the sphere of particles initially distorts is unchanged from that of the sphere. (This assumes that the sphere is small and that the particles start off at rest relative to one another.) Had the sphere surrounded matter, rather than empty space, then there would have been a volume reduction which, according to *Einstein's field equations*, would have been proportional to the total mass enclosed by the sphere (with a slight correction due to pressures). We can see an example of this by taking our sphere to surround the whole earth rather than being situated in empty space (Fig. 9). Now, there is an acceleration inwards all around the

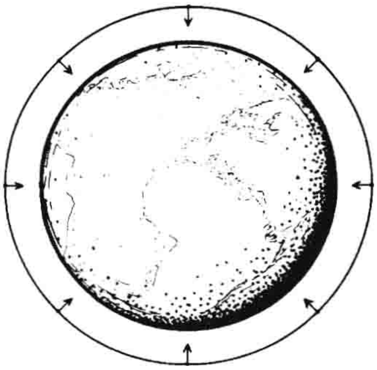


Fig. 9 When the sphere surrounds matter (here the earth), there is a net inwards acceleration. This illustrates the effects of Ricci curvature.

sphere (the normal gravitational acceleration that we experience at the earth's surface), and the volume of the sphere would indeed be reduced. The type of space–time curvature that causes volume reduction is called *Ricci* curvature. Einstein's field equations describe how the Ricci curvature is determined by the energy and momentum densities of matter (and electromagnetic field). There are ten independent components of Ricci curvature, whereas there are twenty components of space–time curvature altogether (the full *Riemann* curvature). The remaining ten components can be collected together into a quantity called the *Weyl* curvature. It is Weyl curvature that describes the distortion, depicted in Fig. 6, due to vacuum space–time curvature. We can think of the Weyl curvature as describing free gravitational field.

The accuracy of Einstein's general relativity is extraordinary. Not only does it inherit all the already remarkable accuracy of Newton's gravitational theory, together with that of Minkowski's space–time (in the absence of gravity), but it achieves much more. It explains the puzzling perihelion advance of Mercury's motion and accurately measured clock-slowing and light-bending effects in gravitational fields. Most impressively of all, it accurately accounts for a slow spiralling inwards of a certain pair of neutron stars in orbit about one another (the binary pulsar PSR 1913+ 16) as an effect due to energy-loss by gravitational waves (a gravitational analogue of Maxwell's electromagnetic waves)—so accurately, in fact, that general relativity can now be regarded as confirmed to an accuracy of about one part in 10^{13}, which makes it, in this sense, the most accurate theory known to science! General relativity is unquestionably a SUPERB theory.

THERMODYNAMICS

The theories that I have been considering up to this point have been concerned with individual bodies and with the nature of the space (–time) in which they move. One might imagine that when large collections of such bodies are considered together, there will be nothing new in principle: it is just a matter of applying the same laws that control individual bodies, albeit in very complicated combinations. Let us restrict attention, for the moment, to Newtonian mechanics. For individual particles, their initial positions and velocities will fix their future motion; the same is true for large collections of particles, and indeed there is a clear sense in which there is nothing in principle that is new when there are very many particles involved.

Of course, with large numbers of bodies, the exact calculations could get prohibitively complicated, but one would expect that approximate procedures could be developed in order to provide an overall picture of what the implications of the dynamical laws would be, even though it would not be possible to keep track of individual particles. The initial positions and velocities of such particles would, in any case, not normally be known in practice, and only a comparatively small number of averaged overall parameters (such as temperature or pressure, for example) might be accessible to observation. Thus, such approximate mathematical procedures would be very valuable—indeed, essential—for the implications of the dynamical laws as applied to individual particles to be understood at the macroscopic level of everyday experience.

But is there anything *physically* new that is needed here; or is it just a matter of mathematical technique? I should make it clear that there has indeed been a vast amount of very valuable mathematical technique (a subject referred to as statistical mechanics) that has been developed to handle such matters, but what is not so readily appreciated is that there *is* actually a new *physical* principle that has crept in, almost unnoticed, under the guise of the *second law of thermodynamics*. This law asserts that the *entropy* of a closed system increases with time (or perhaps remains a constant —in the case of a reversible system). Incidentally, the *first* law of thermodynamics is a statement that the total energy of a closed system remains constant, where the energy content of heat is taken into account. (The subject of thermodynamics arose, mainly in the nineteenth century, from the work of Carnot, Maxwell, Kelvin, Boltzmann, and others, and it developed, in this century, via the work of Gibbs and Einstein, into the subject of statistical mechanics.)

What is entropy? It is best understood in terms of the concept of *phase space*. Phase space is a mathematical space, normally of very large dimension, each of whose points represents the entire state of a physical system at one time, the coordinates of the phase-space point encoding the positions and momenta of all of the various particles that constitute the system (Fig. 10). The dynamical evolution of a system is then described as just the motion of a single phase-space point. To define the (Boltzmann) entropy of a system, we must perform what is called a *coarse-graining* of the phase space (Fig. 11), which means dividing it up into large compartments, where the points in any one compartment are to be taken as representing states that are *indistinguishable* from one another as regards their macroscopic properties. The entropy of a system is then the *logarithm* of the phase-space *volume* (taken in appropriate units) of the compartment to which the point representing that system belongs.

To understand the rough meaning of the second law, think of a system that starts off in some particular configuration of reasonably small entropy and then proceeds in such a way that its entropy increases. For example, consider a glass of water perched on the edge of a table (see Fig. 12). If the glass is nudged, then it is likely to fall to the floor, shattering and spilling its contents irretrievably. In terms of the ordinary Newtonian dynamics of the constituent particles, everything is reversible, and

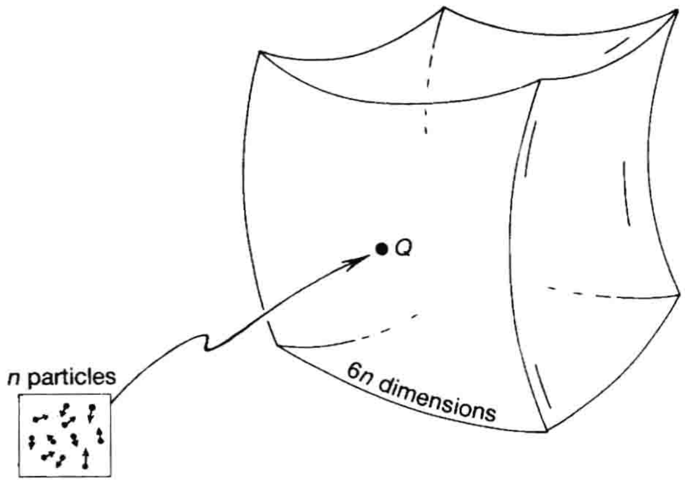


Fig. 10 Phase space: a single point *Q* of phase space represents the entire state of some physical system, including the instantaneous motions of all its parts.

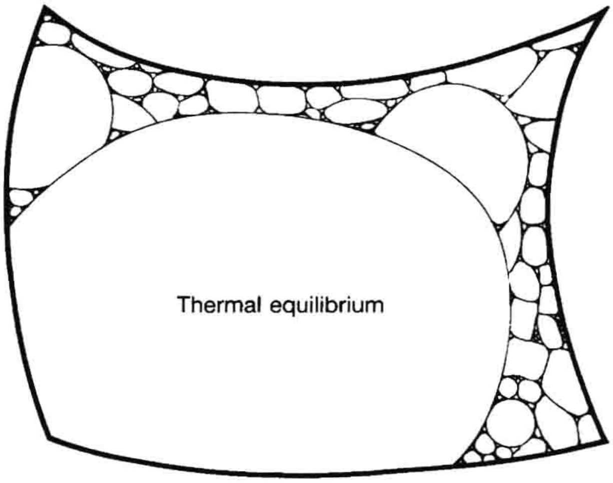


Fig. 11 A coarse-graining of phase space into regions corresponding to states that are macroscopically indistinguishable from one another. The *entropy* is proportional to the logarithm of the phase-space volume.

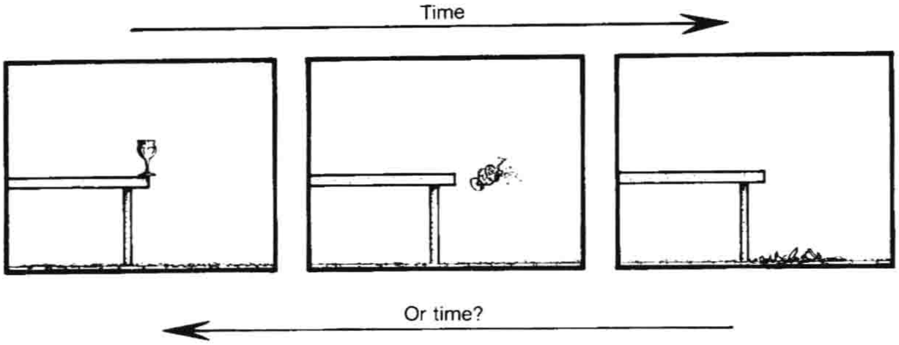


Fig. 12 The laws of mechanics are time reversible; yet the time-ordering of such a scene from the right frame to the left is something that is never experienced, whereas that from the left frame to the right would be commonplace.

there would be nothing against the whole process occurring, instead, in the reverse temporal direction, with little specks of water scattered about the floor collecting together into a glass, which simultaneously assembles itself from innumerable fragments, jumping delicately up in order to balance itself just on the edge of the table. The first progression, however, would be commonplace whilst the second would never occur in practice. What distinguishes the two, from the point of view of thermodynamics, is that the first is in accordance with the second law, whereas the second grossly violates it. The shattered glass and dispersed water droplets provide a state of considerably higher entropy than does that of the glass of water perched on the table. A shattering glass represents an increase in entropy, whereas a self-assembling one would represent a decrease in entropy.

In phase-space terms, we can understand the second law in the following way. Our phase-space point starts in a relatively tiny compartment, but as the point moves, it tends to enter larger and larger compartments as time progresses, see Fig. 13. This is natural, for the larger compartments are liable to have volumes that are utterly enormous compared with the smaller ones, so that once the point has entered a larger one there will, in effect, be no chance that it will find the smaller one again. Explained in this

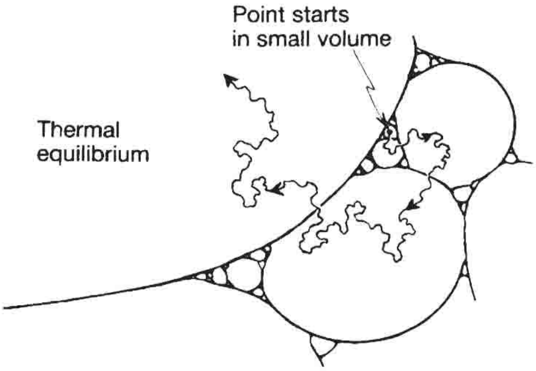


Fig. 13 The second law of thermodynamics in action: as time evolves, the phase-space point enters compartments of larger and larger volume. Consequently the entropy continually increases.

way, the second law seems almost obvious, and it is hard to see why there should be any new physical principle involved here.

However, this 'obviousness' is somewhat deceptive. The laws of Newtonian mechanics are symmetrical with respect to a reversal of the direction of time, yet the second law itself is grossly time *a*symmetric. How can we have deduced a time-asymmetrical law simply from time-symmetrical ones? In fact, we have not. We have made the implicit assumption that it is all right to *start* a system off in a state of low entropy, and then deduce its probable future evolution assuming that there are no low-entropy constraints in the future. If we attempted to use a similar method of reasoning for the most probable *past* behaviour for a system, merely being given its present state, then we would get nonsense. (For example, we would deduce that the most likely way for a glass of water to have got to a state in which its sits perched on the edge of a table is that it spontaneously assembled itself from broken pieces and scattered water droplets, and jumped on to the table by itself!) The difference is that, in the universe that we happen to inhabit, there *was* an enormous low-entropy constraint in the past, whereas there do not appear to be any such constraints in the future.

What was it that happened in the past? I do not want to go into the full discussion of this here, but the upshot of it all is that there must, indeed, have been an absolutely stupendous constraint at the very beginning of time—at an event referred to as the *big bang* origin of the universe. According to modern cosmology, the entire universe was created some 10^{10} years ago in a gigantic explosion. There is a very impressive theory, called the standard model of the hot big bang, that belongs to, at least, the USEFUL category of physical theories. It has many different kinds of confirmation, ranging from the large-scale expansion and uniformity of the universe to detailed information about the nature and proportions of different kinds of atomic elements and particles. (I am almost inclined, myself, to place this theory in the SUPERB category, but perhaps one should be more cautious than this. On the other hand, there are variants and extensions of the standard model that have become popular in recent years, such as 'the inflationary scenario'. But these, it seems to me, cannot yet rise above the category TENTATIVE, in view of the lack of any detailed confirmations.)

How does the standard picture of the big bang serve to constrain the initial universe? I should explain that there are really three distinct classes of standard model, depending upon the spatial geometry that results from the big bang. In one of these models the spatial geometry is Euclidean (referred to as the $k = 0$ model), while in the other two it is either Lobachevskian ($k = -1$) or spherical ($k = +1$). The first two of these models ($k = 0$ or -1) continue to expand indefinitely, whereas the third ($k = +1$) finally recollapses to an event known as the *big crunch*. If anything, present observational evidence favours $k = -1$, but there are very considerable uncertainties, and it is possible that $k=0$ or $+1$ might eventually turn out to be more appropriate. For convenience, I shall consider only the $k = +1$ case here, for which the arguments are somewhat clearer than in the other two cases. In Fig. 14, I have given space–time pictures of the early universe and of its entire history in the idealized standard $k = +1$ case. I have suppressed two spatial dimensions, so that the entire spatial universe appears as a horizontal circle (really representing a three-dimensional spherical surface—the surface of a sphere in four-dimensions). The model is idealized because I have ignored all irregularities in the space–time geometry that should be present (according to Einstein's field equations) due to irregularities in the actual

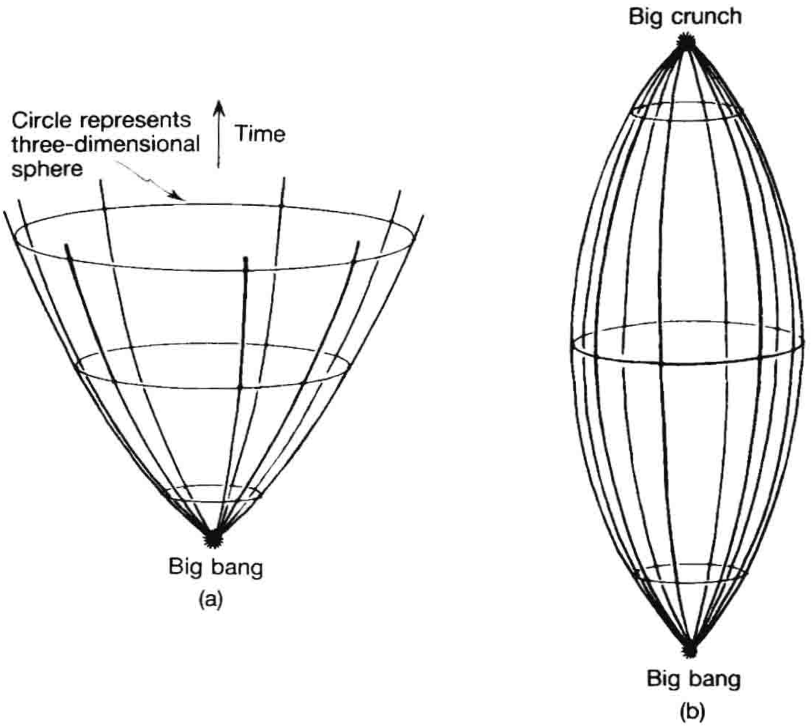


Fig. 14 (a) Space–time picture of an expanding universe with spherical spatial sections (only one space dimension depicted). (b) Eventually this universe recollapses to a final big crunch.

matter distribution. If we include such irregularities, then the picture would be more like that of Fig. 15, in which these irregularities continually increase as time progresses. The final big crunch which can be thought of as the congealing of a large number of black holes) has an immensely more complicated structure than the big bang. This corresponds to the fact that the entropy of the big crunch would be enormously large, whereas that of the big bang was comparatively very tiny, in accordance with the second law. The geometry of the big bang must have been constrained in such a way that corresponds to this low entropy. In effect, this means that the gravitational degrees of freedom that correspond to possible irregularities in the geometry, and which were in principle

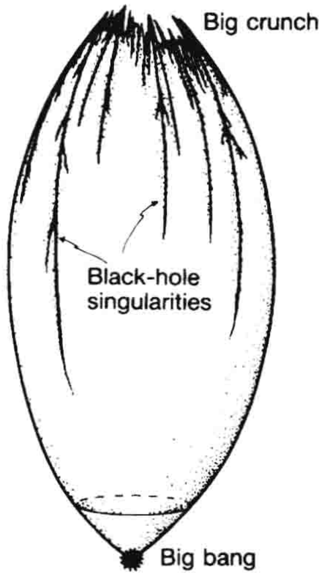


Fig. 15 The entire history of a closed universe starting from a low-entropy big bang, with vanishing Weyl curvature, and ending with a high-entrophy big crunch (representing the congealing of many black holes) where the Weyl curvature diverges to infinity.

available to the universe in its early stage, were, for some reason, not actually 'excited' in the big bang as it actually occurred. This remarkable fact has given us the second law of thermodynamics.

I should make it clear to the reader that in the above descriptions I have presented the second law of thermodynamics in a somewhat personal way. Most other physicists would probably not have emphasized cosmology as I have done. Moreover, there are certainly many important aspects of the second law that I have not considered. However, it seems to me that we must turn to cosmology if we are to gain some genuine understanding of a powerful new physical principle, of a time-asymmetrical nature, that must indeed be adjoined to the physical laws that we already understand. Part of this principle must manifest itself in the *boundary condition* for the universe that was the big bang—although there may well be other aspects of such a principle, to do with (as I believe) missing ingredients in relation to quantum theory (and which could conceivably even provide some apparently 'organizational' aspects to structures at a biological or crystallographic level, perhaps in accordance with ideas of some other workers in this field, such as Prigogine).

The big bang is an example of what is called a *space–time singularity*, where the space–time curvature becomes infinite and the equations of general relativity reach their limits of applicability. Other examples of space–time singularities occur in gravitational collapse: in black holes or the big crunch. In order to understand, in physical terms, what is going on in space-time singularities, it will be necessary to obtain a fundamental new physical theory which correctly relates general relatively with quantum theory: *quantum gravity*. Such a theory is at present lacking, but it seems clear to me that one of its features must be a *time asymmetry*, in order that the structure of singularities will differ profoundly depending upon whether they are of *past-type* (the big bang) or of *future-type* (in black holes or the big crunch). The past-type singularities appear to be enormously constrained (low entropy)—apparently in such a way that the Weyl part of the space-time curvature goes to zero (or is at least comparatively small, in relation to the Ricci part), whereas the future-type singularities are not so constrained. The enormity of such a constraint can be seen by an example: if the universe (assuming $k = +1$) contains of the order of 10^{80} particles, then the region of phase space that it would be constrained to at the big bang would be only about one part in $10^{10^{123}}$ of the entire phase-space volume! (This corresponds to an entropy difference of 10^{123}; recall the logarithm in the definition of entropy. The value depends upon an expression for black-hole entropy due to Bekenstein and Hawking.)

When I refer to thermodynamics as a SUPERB theory, I do not mean to be taking figures like those just given as a measure of its accuracy, since in my discussion I have taken things considerably away from what is standard in thermodynamic theory—which does not normally involve considerations of cosmology. Nevertheless, there is indeed something extraordinarily precise about the second law. This is despite the impression of imprecision involved in the concept of entropy, an impression that is in many respects justified in view of the arbitrariness in the concept of coarse-graining, etc. For reasons such as this, I had initially felt reluctant to place thermodynamics in the SUPERB category—or, indeed, to think of it as a physical theory at all in the sense of the others. On reflection, however, I have come to the conclusion that its characterization as a SUPERB physical theory is amply justified, in view of the extraordinary degree of organization that it implies for the structure of the universe in the past.

QUANTUM MECHANICS

The most enigmatic of the SUPERB theories is undoubtedly quantum mechanics, which arose in the first quarter of this century from the combined efforts of many outstanding scientists: particularly Planck, Einstein, Bohr, Heisenberg, Schrödinger, de Broglie, Born, Jordan, Pauli, and Dirac. If we take into account the breadth of applicability of this theory and its far-reaching nature, then is it also the most accurate of the SUPERB theories (where I mean this accuracy to be taken in a different sense from the precision exhibited in a single phenomenon, in the case of general relativity, or the somewhat obscure sense of the precision involved in thermodynamics) The enigmatic nature of this theory lies in the combination of this extraordinary precision when the theory is applied to phenomena that, in a certain sense, are on a 'small scale'—I shall call this scale the *quantum level*—and a bewildering absurdity if one tires to apply the theory in the same way at the macroscopic level of everyday experience—called the *classical level*. Almost as bewildering is the multitude of different attitudes that are adopted by physicists and philosophers in order to come to terms with, or perhaps explain away, the apparent (or real?) absurdities. It is inevitable that, in a short account like this, I shall display a considerable personal bias. I make no apology for this, but the reader is hereby warned that not everyone views the subject quite as I do. What is *not* controversial, however, is the *mathematical* formalism of the theory, and how this formalism is applied in practice, so let me start by giving a very brief account of what we shall need of this.

There is a very basic principle in quantum mechanics known as the *superposition principle*, which asserts that if two alternative things $\mathbb{A}$ and $\mathbb{B}$ might happen, then weighted combinations $w\mathbb{A}+z\mathbb{B}$ can also happen. What does that mean? It may be vaguely helpful if just for the moment we try to think of w and z as *probabilities*. Then $w\mathbb{A}+z\mathbb{B}$ would be interpreted as some kind of probability-weighted assertion that either $\mathbb{A}$ or $\mathbb{B}$ might be happening, but we do not know which, and the relative probability of $\mathbb{A}$ to $\mathbb{B}$ is the ratio of the numbers w to z. This might have been fine had w and z been ordinary positive real numbers, but—and this is the shock—w and z are *complex* numbers!

I should remind the reader what a complex number is. It is a number constructed from ordinary real numbers and the *square root of minus one*. More precisely, it is a number that can be written as $x+\mathrm{i}y$, where x and y are real numbers and where $\mathrm{i}^2=-1$. This may seem very mysterious, but for the moment, just think of this as a bit of mathematics. I shall try to illucidate the physical meaning of all this in a moment. As mathematical objects, complex numbers make perfectly good sense. You can add them or multiply them out: if $w = u + \mathrm{i}v$ and $z = x + \mathrm{i}y$, then $w + z = (u + x) + \mathrm{i}(v + y)$ and $wz = (ux - vy) + \mathrm{i}(vx + uy)$, where u, v, x, and y are real. It is also very helpful to use a geometrical representation of complex numbers as points in a Euclidean plane (often referred to as the *Argand plane*; see Fig. 16), where the complex

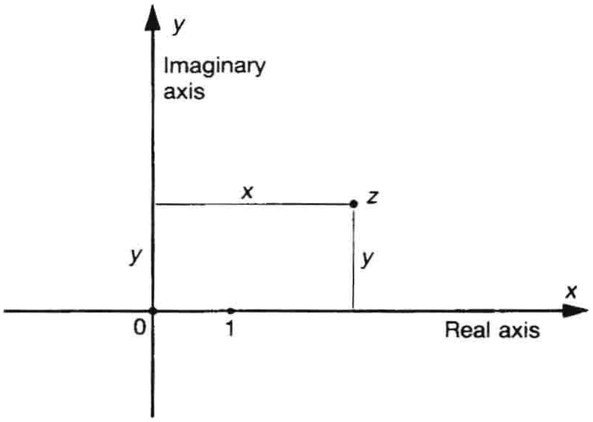


Fig. 16 The Argand plane, depicting a complex number $z = x + \mathrm{i}y$.

number $z = x + iy$ is represented as the point with Cartesian coordinates (x,y). The rules for adding and multiplying complex numbers have simple geometrical interpretations in terms of the Argand plane.

Let us now return to quantum mechanics? How are we to interpret the superposition $w\mathbb{A} + z\mathbb{B}$? It is certainly *not* a probability-weighted pair of alternatives after all, since a probability cannot be a general complex number. Instead, we must think of it as just another physical alternative, different from either $\mathbb{A}$ or $\mathbb{B}$. At least this is how things seem to behave at the quantum level. As soon as the effects of $\mathbb{A}$ and of $\mathbb{B}$ can independently get magnified to the classical level, then we do *not* allow the possibility of $w\mathbb{A} + z\mathbb{B}$, but we find that *either* $\mathbb{A}$ *or* $\mathbb{B}$ occurs with indeed a probability weighting, but the relative probabilities of $\mathbb{A}$ and $\mathbb{B}$ are not w and z, but their *squared moduli* $|w|^2$ and $|z|^2$. (The squared modulus of a complex number z is the square of the distance that the point representing it in the Argand plane lies from the origin, i.e. it is $|z|^2 = x^2 + y^2$, where $z = x + iy$, with x and y real.)

The procedure whereby the quantum state $w\mathbb{A} + z\mathbb{B}$ appears suddenly to get converted into either the state $\mathbb{A}$ or the state $\mathbb{B}$, with relative probability weightings $|w|^2$ and $|z|^2$, whenever the effects of $\mathbb{A}$ and $\mathbb{B}$ get amplified to the classical level, is called *reduction of the state vector* (or collapse of the wavefunction), and I denote it by the letter **R**. (Strictly speaking, this happens only for states $\mathbb{A}$ and $\mathbb{B}$ that are what are called *orthogonal*, but I shall not try to define precisely what this means. Think of $\mathbb{A}$ and $\mathbb{B}$ simply as being appropriately *independent* of one another.) For a system that remains entirely at the quantum level, however, the process **R** seems never to occur. Instead, there is a *deterministic* procedure, that I denote by the letter **U**, giving the evolution of the quantum state, called *unitary evolution*, which is governed by a precise (partial) differential equation called the *Schrödinger equation*. The procedure **U** differs from **R** in almost every conceivable respect: **U** is continuous, deterministic, local (in configuration space), and time-symmetric, whereas **R** is discontinuous, non-deterministic, non-local, and time-asymmetric. Moreover, **U** is *linear*, in the sense that the **U**-evolution of the state $w\mathbb{A} + z\mathbb{B}$ is precisely the same as the complex-weighted superposition of the independent

U-evolutions of the state $\mathbb{A}$ and the state $\mathbb{B}$ separately, where the complex weighting factors *w* and *z* remain *constant*. This is certainly not true for **R**.

Many physicists are of the opinion that the process **R** should not be considered as a real physical process, but it arises as some kind of 'illusion' or approximation, when a system becomes large or complicated, or when it becomes inextricably interconnected with the physical environment. For my own part, I cannot see how this alone could be the case. The procedures **U** and **R** are just too mathematically different from one another; in particular, I cannot see how two superposed states can ever get themselves *un*superposed, merely by evolving according to the procedure **U**, nor do I see how the time-asymmetry inherent in **R** can arise merely through the action of **U**. Other physicists (or, quite often, the same physicists as before) try to take the line that it is **U** that is the illusion, and there is actually no 'reality' at all at the quantum level. Instead, they may claim to be driven to an entirely 'subjective' view of the quantum world (although this is not the usual sense of the word 'subjective', as far as I can make out) and often they would simply regard the mathematical formalism as being merely a convenient procedure for calculating various probabilities for the different results of 'measurements' that might be performed on a system. I have great difficulties with such viewpoints also, since I do not see how we can dispense with physical reality, even at the quantum level. Moreover, the mathematical procedures for describing the quantum level are so precise and accurate that I do not see how they can fail to be reflecting some kind of precise reality that is indeed taking place at the level of quantum phenomena.

In order to clarify the mathematical procedures, it will be helpful to take a look at a simple example of a quantum system. Let us consider a single particle of 'spin one-half', say an electron or proton. Its state of spin can be characterized by a direction in space about which it may be spinning in a right-handed or left-handed sense. The usual convention is to take the right-handed sense as counting as positive and the left-handed one as negative, so we take the left-handed particle as spinning (positively) about the *opposite* direction in space. Now, for a spin-one-half particle we can take all the states of spin to be superpositions of (states of

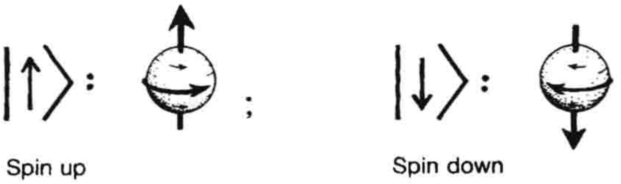


Fig. 17 Two independent (orthogonal) states for the spin states of a spin one-half particle, e.g. an electron, are *spin up* (right-handed about the upward vertical) and *spin down* (right-handed about the downward vertical). All other spin stages are expressible as combinations of these.

spin about) just two opposite directions, say up, written$|\uparrow>$, and down, written $|\downarrow>$ (see Fig. 17.) The general state of spin, say $|\nearrow>$, will be some superposition of these two states: $|\nearrow>=w|\uparrow>+z|\downarrow>$. There is nothing special about the directions 'up' and 'down', so there is nothing special about our original states $|\uparrow>$ and $|\downarrow>$ (except that they must be orthogonal to each other, which in this situation means *opposite* directions in space). We could equally have chosen any other pair of opposite directions to start with, say 'right' $|\rightarrow>$ and 'left' $|\leftarrow>$, and then expressed the general state of spin (and, in particular, 'up' or 'down') in terms of these other two. This illustrates the fact that, at the quantum level, superposed states are no less (and no more) real than the original states that they are superpositions of. Another point is also illustrated. This is that the complex numbers that appear in these superpositions, though mysterious in their interpretations in relation to probability, make good *geometrical* sense. The geometrical space of possible ratios $w{:}z$ of pairs of complex numbers w and z is a *sphere* (referred to as the *Riemann* sphere), and this sphere can be directly related to the Argand plane that we considered earlier (see Fig. 18). Here, the complex number $q=z/w$ corresponds to a point on the Argand plane (or to infinity, if $w=0$), and we can refer this to the sphere by projection from the south pole (a correspondence referred to as *stereographic* projection). The point that q corresponds to on the sphere gives the direction of spin for the quantum state $|\nearrow>=w|\uparrow>+z|\downarrow>$

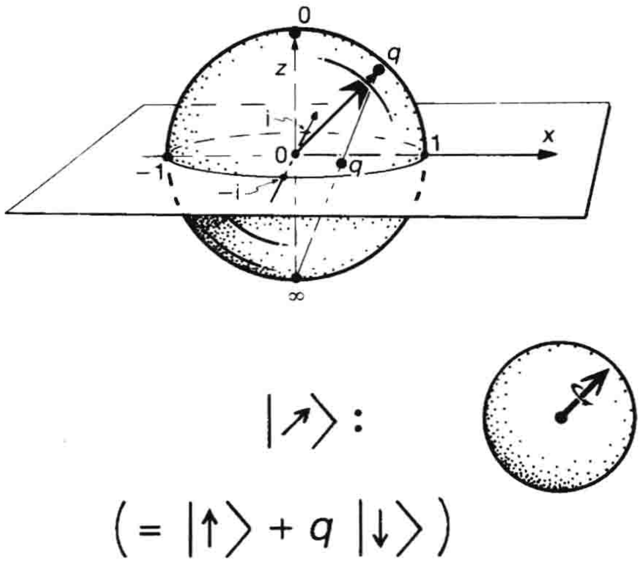


Fig. 18 The Riemann sphere, here represented as the space of physically distinct spin states of a spin one-half particle. The sphere is projected stereographically from its south pole (∞) to the Argand plane through its equator.

(which is physically the same as $w^{-1}|\nearrow\rangle = |\uparrow\rangle + q|\downarrow\rangle$, since only the *ratios* of the coefficients w and z are regarded as physically significant, according to the theory).

If we now wish to 'make a measurement' of the spin, we select a direction in space and orient our measuring apparatus in that direction. Let $\mathbb{A}$ represent the state of spin in that direction and $\mathbb{B}$, the state of spin in the diametrically opposite direction. The apparatus serves to amplify the effects of $\mathbb{A}$ and $\mathbb{B}$ to the classical level, so we must now regard all possible states of spin as superpositions of *these* two, whatever the original state of spin might have been. The result of our measurement must be either $\mathbb{A}$ or $\mathbb{B}$, with respective probabilities in fact given by $\cos^2\theta$ and $\sin^2\theta$, where the original state of spin of our particle was tilted at an angle θ to the chosen direction. The state of spin must now 'jump' to either $\mathbb{A}$ or $\mathbb{B}$, with these respective probabilities, in accordance with the procedure **R**, whereas if we had chosen *not* to make a measurement, the state of spin could have been kept at the quantum level, and it would then have evolved, perhaps precessing in some way, or remaining constant, according to the procedure **U**, and this evolution could be checked by performing some other measurement at some later time.

An important feature of quantum mechanics is the way that systems involving several particles must be treated. For example, for a system consisting of just two particles, say an electron and a proton, the state $\mathbb{A}$ might describe the electron as being at some particular point E and the proton, at some point P, while the state $\mathbb{B}$ might describe the electron as being at F and the proton at Q. The superposition principle asserts that $\mathbb{A} + \mathbb{B}$ is also a possible state. This is an example of what is called an *entangled* (or correlated) state, and it exhibits a peculiar property characteristic of such states. Suppose that we decide to measure the position of the electron (i.e. we set up some experimental situation in which the respective effects of the electron being at E or F are magnified to macroscopically distinguishable outcomes). Then we find that the whole state must jump to $\mathbb{A}$ if the electron is found to be at E and to $\mathbb{B}$ if the electron is found to be at F. The remarkable thing about such a measurement of an entangled state is that the measurement of the *electron* seems to cause the *proton's* state to jump simultaneously with it: in the first case, the proton ending up at the point P and in the second, at Q. This would be the case no matter how distant the proton might be from the electron. There would appear to be some mysterious action of the measurement on the proton, even though it is only the electron that is being measured. This action seems to have to occur *instantaneously* even though the proton could be many metres, or perhaps light-years away from the measuring apparatus. This is what Einstein, in expressing his deep unease with the physical implications of quantum theory, referred to as 'spooky action at a distance'. On the face of it there would appear to be a contradiction with the principles of Einstein's relativity, since the very concept of what 'instantaneous' means is dependent upon the motion of the 'observer' in whose reference frame the physical description is being made. In fact there is no conflict with the *observational* aspects of relativity; in particular it is not possible to send a message faster than light using this 'spooky action', although there is a distinct conflict with the *spirit* and with the space–time descriptions of relativity theory. There are serious problems that are still not completely resolved in making the principles of special relativity and quantum mechanics consistent with one another. (We shall be touching on a little of this in the next section.) The problem of combining the principles of *general* relativity and quantum mechanics is considerably more serious, and remains one of the deep stumbling blocks to fundamental progress in theoretical physics.

I should clarify one point about Einstein's 'spooky action'. One can envisage an apparently similar situation that can occur with ordinary classical bodies. Imagine a box which is known to contain one white ball and one black ball. Suppose that the balls are taken out and removed to two opposite corners of the room, without either being looked at. If one ball is examined and found to be white (like the electron, above, being found to be at the point E) then the other ball must be black (like the proton being consequently at P). If, on the other hand, the first ball is found to be in the state 'black' (like the electron found at F), then the second ball's uncertain state appears to jump instantaneously to 'white' (proton found at Q). But no one would attribute the sudden change of the second ball's 'uncertain' state to being 'black with certainty' or to being 'white with certainty', to some mysterious non-local 'spooky action' travelling to it instantaneously from the first ball the moment that first ball is examined. The point is, however, that the quantum-mechanical situation often can*not* be treated in the same way as the balls just considered, where in that case there was an underlying 'reality' about where the white and black balls actually were. In quantum mechanics, one has to treat an entangled state as a single entity and not as a pair of subsystems that behave independently of one another (like the black and white balls) when they are widely enough separated from one another.

An example (suggested by Bohm, illustrating earlier ideas due to Einstein, Rosen, and Podolsky—and often referred to as the 'EPR paradox') is an entangled system consisting of two spin one-half particles (say an electron and a proton) that are known to start in a state for which the total spin is zero. When they are separated, the spin of the electron must be opposite to that of the proton, though the spin direction of either may not be defined on its own. If some direction in space is selected, and the electron's state of spin is measured in that direction, then it must (instantaneously) turn out that the state of spin of the proton is in the opposite direction. If one could measure the actual direction of the proton's state of spin then it would be possible to use the orientation of the electron-measuring apparatus as a signalling device that could transmit messages faster than light (since the proton's spin-axis is apparently instantaneously fixed to be in the selected direction). However, all that the rules of quantum mechanics allow is that that probability values are given to the two possible spin values (i.e. right-handed or left-handed) about a *chosen* direction for the measurement of the proton's spin. It turns out that these probabilities are such that (a) no signal can indeed be sent in this way, and (b) *no* description can in fact be given where the electron and proton are considered as independent physical objects (like the black and white balls considered above)—as follows from a famous theorem due to Bell. Certain careful and ingenious experiments of this general nature have been carried out over the past twenty years, and have demonstrated, via Bell's theorem, the inadequacy of a 'local' description of physical reality, in which the electron and photon in the above description could be considered as independent objects. The most impressive of these experiments was performed by Aspect and his co-workers in Paris in the early 1980s. The experiment used photons, rather than the electron–proton pair of my description above, but the principle is the same.

There are many other respects in which quantum mechanics remains an enigmatic and baffling theory. Most of these have to do with the **U-R** conflict referred to above, and it is fair to say that there are fundamental aspects of the interpretation (at least) of the theory that are unresolved. It seems that we must use one deterministic theory at the quantum level of phenomena—where **U** holds strong—and another quite different deterministic theory (classical physics) at the classical level. When effects are magnified from the quantum to the classical level, we must bring in the *in*deterministic procedure **R**. There are some physicists (probably only a small minority) who believe that **U** holds at all levels, but that when a 'measurement' takes place the observer's conscious state becomes entangled with the result of that measurement, so that his being consciously aware of one outcome of an experiment is entangled with that particular outcome, while his awareness of another outcome is entangled with that other outcome; yet both states of awareness and both outcomes must coexist in the 'actual' state of the universe. This is the so-called 'many-worlds' view (although a much larger number of 'pragmatic' physicists would claim to be able to follow this route without reaching its 'many-worlds' logical conclusion). Another (probably also unpopular) view is that the rules of quantum theory must be changed whenever consciousness enters the picture. My own view (no doubt also unpopular) is that the rules of quantum theory must indeed be changed at some level, but this change will arise because of the effects of quantum gravity (which implies that, roughly speaking, the process **R** comes in when the superposed states of a quantum system get entangled with as much mass as the so-called Planck mass, which is abut 10^{-5} grams) and that it is in relation to the time assymetry that must be present in quantum gravity that the time-asymmetric aspects of the **R** procedure will eventually be understood. In any case, even the standard theory is unclear about what it is that actually constitutes a 'measurement', so that, for whatever reason, the process **R** comes into play. Problems remain, no matter what viewpoint one adopts.

Be that as it may, quantum mechanics is undoubtedly a SUPERB theory. Its applicability to physical behaviour at the quantum level remains supreme, and no phenomena are known that contradict the theory. Moreover, it explains many things, both familiar and novel that seem incomprehensible on the basis of classical ideas (such as spectral lines, chemical bonding, the colours of heated substance, solidity, freezing and boiling, the accuracy of heredity, lasers, superconductivity, superfluidity, etc.) always with complete agreement between theory and observation—to the extent that the theory can, at present, be actually worked out.

QUANTUM ELECTRODYNAMICS

The subject of *quantum field theory* is the application of the procedures of quantum mechanics to a (relativistic) classical field theory—such as Maxwell's electrodynamics. We say that it is the theory obtained by *quantizing* that field theory. Alternatively, we may view quantum field theory as the subject that has arisen from attempts to make quantum theory consistent with special relativity. In the context of quantum theory, 'fields' and 'systems of identical particles' turn out to be the same thing, so that quantizing Maxwell's field, for example, is the same as quantizing a system of relativistic particles of the appropriate kind—in this case 'photons', which are massless particles of unit spin. I have chosen to separate quantum field theory from ordinary (non-relativistic) quantum theory, as described in the previous section, because quantum field theory brings in a number of new ingredients, and these can give rise to some very serious problems. The nature of these new problems is that if one strictly applies the quantum-mechanical rules to an interacting relativistic field, then one simply gets the answer 'infinity', rather than something sensible. In a general way, this comes about because a field theory involves an infinite number of degrees of freedom, so the quantum superpositions that have to be considered in a process are much more extensive than they are in a theory with a finite number of particles. These infinities have to be eliminated from the theory before it can be used. This turns out to be possible for certain types of quantum field theories—called *renormalizable theories*—but the procedures are somewhat artificial, going strictly outside what quantum theory tells us to do, and they involve the loss of certain kinds of information from the theory. For this reason, quantum field theory is, strictly speaking, a different theory from ordinary quantum theory, and its status as a SUPERB theory, or otherwise, must be considered separately.

The most successful of the quantum field theories is undoubtedly *quantum electrodynamics* (or QED) which involves not only the quantized electromagnetic field—i.e. *photons*—but also *electrons*, using Dirac's famous equation for the electron that he introduced in 1931. QED emerged from work of Jordan, Heisenberg, and Pauli in the mid-1920s; it was formulated by Dirac in 1926–34, but had to wait until 1947–9 to be made workable by the elimination of infinities, according to the process of *renormalization*, as was done by Bethe, Schwinger, Feynman, and Tomonaga. Feynman also introduced a number of conceptual and technical innovations that have been very influential. One of these was the *path integral* (based partly on an earlier idea due to Dirac) according to which all possible space–time *histories* must be superposed—including histories in which particles are allowed to wander backwards and forwards in time as well as space (see Fig. 19). The interpretation of an electron, when it wanders backwards in time, is that then it is a *positron*—i.e. an *anti*-electron. As the electron's space–time path bends around from being future-pointing to past-pointing, this is interpreted as the annihilation of the electron by a positron; and when the path bends back again from being past-pointing to future-pointing, there is creation of a new electron, in association with this positron.

In fact, the physical existance of positrons was a prediction of Dirac's original theory of electrons, and it was soon confirmed in observation (by Anderson in 1932). Moreover, it is a general consequence of the principles of quantum field theory that *every* particle must have its corresponding *antiparticle* (sometimes, as in the case of a photon, a particle of the same kind as the original one). Any massive particle can annihilate its antiparticle to form energy, and such a pair can be created out of energy. It is a

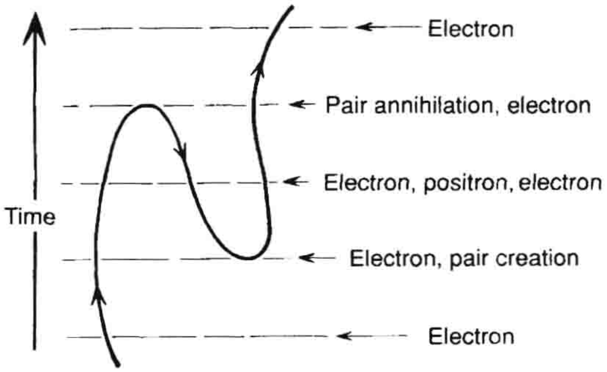


Fig. 19 In a Feynman path integral, all possible classical histories, for the type of particle in question (here an electron), must be superposed. The classical history can involve the particle travelling 'backwards in time' which is interpreted as the particle's antiparticle (here a positron) travelling forwards in time. One such contribution to the path integral is depicted here.

characteristic feature of quantum field theory that the number of particles of each different kind, in a physical system, need not be a constant; nor need this number even be definite, since superpositions of states with different numbers of particles and antiparticles are allowed.

Another influential (and related) innovation due to Feynman (and Dyson) was what are now called *Feynman diagrams*. These are graphical representations of certain mathematical expressions (integrals—which, when taken at face value, are unfortunately usually infinite) that can be used to calculate virtually anything that is needed in quantum field theory. What makes Feynman diagrams conceptually very convenient is that they give us pictures of the various physical processes that can contribute, in superposition, to the total effect. In Fig. 20, I have drawn some of the Feynman diagrams for a situation in which an electron and a positron annihilate to produce two photons. The diagrams shown

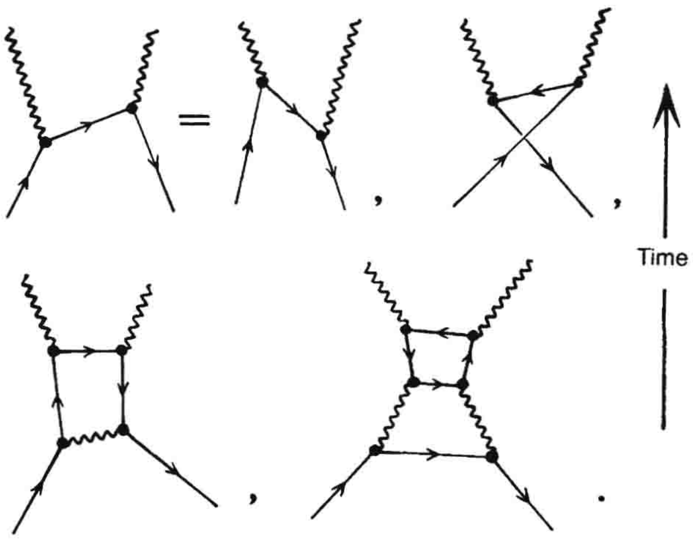


Fig. 20 Various Feynman diagrams are depicted, representing processes that contribute, in superposition, to the total effect when an electron and a positron annihilate to produce two photons. (Electrons and positrons are represented as straight lines with arrows; photons are represented as wavy lines.)

are all concerned with QED, where only electrons, positrons, and photons are involved. Other quantum field theories would involve further particles, such as mesons, or quarks, gluons, etc.

QED exhibits an extraordinary precision, in the comparatively small number of situations for which the theory can be worked through and compared with accurate experiment. In particular, the value of the magnetic moment of an electron is calculated to be 1.00115965246 (in appropriate units, with an allowance for an error of about 20 in the last two digits). This may be compared with a recent experimental value, which is 1.00115965221 (with a possible error of about 4 in the last digit). (Electrons behave like tiny magnets of spinning electric charge, and the term 'magnetic moment' refers to the strength of this tiny magnet.) As Feynman has pointed out, this kind of accuracy could determine the distance between New York and Los Angeles to within the width of a human hair. There are a few other results, of comparable accuracy. Quantum electrodynamics is indeed a SUPERB theory.

THE USEFUL QUANTUM FIELD THEORIES

No quantum field theory other than QED can match this kind of accuracy. However there are two quantum field theories that , it seems to me, come clearly into the USEFUL category. The first of these is called *electroweak* theory, and the second is referred to as *quantum chromodynamics*, or QCD. The first, put forward by Salam, Ward, Glashow, and Weinberg, in about 1967, is a theory of the *weak* interactions that are responsible for radioactive decay, combining them, in an appropriate way, with QED. The second, QCD, is a theory of the *strong* interactions that bind protons and neutrons together to form atomic nuclei; it is also a theory of the very structure of the strongly interacting particles themselves (protons, neutrons, mesons, etc.). These particles are regarded as being built up, according to a scheme due to Gell Mann and Zweig, from more elementary ingredients, called *quarks* (and gluons), although these ingredients are never seen on their own. Both electroweak theory and QCD are what are called *gauge* *theories*, and are based on a generalization of Maxwell's electrodynamics called *Yang–Mills* theory that was introduced by Yang and Mills in 1954, although, unknown to them, the theory had also been studied a good deal earlier by Pauli (and by Shaw). In order to be applicable to the (strong or weak) interactions of actual physical massive particles, Yang–Mills theory has to be modified from the original version of the theory (where, like the photon, the mediating particles need to be massless). The necessary modifications arose out of a prescription due to Higgs (and Kibble). An ingredient that is incorporated into electroweak theory is the parity-violating effect (mirror-reflection asymmetry) that was observed in weak interactions (beta-decay) by Wu and her collaborators after having been predicted by Yang and Lee. According to the theory of QCD, there is an (unobservable) quantum property referred to as 'colour' (nothing to do with ordinary colours), from which QCD gets its name. There are certain interconnections and consistency requirements between electroweak theory and QCD, and the combination of the two theories is sometimes referred to as the *standard model* of particle physics.

Although there are no observed discrepancies with either electroweak theory or QCD, as far as I know, there is nothing like the very precise agreement with experiment that can be seen in the case of QED. It is perhaps fair to say that there is about *one* decimal place of numerical agreement between the predictions of the standard model and observation. It may be that this comparatively poor level of agreement is merely a reflection of the calculational difficulties that are inherent in the theory. On the other hand, it is possible that these calculational difficulties are simply serving to obscure some essential defect that may be present. Also, it must be said that there is a considerable untidiness in the standard model, and there are many apparently arbitrary parameters involved (masses of basic constituent particles, strengths of basic interactions, etc.). It seems clear to me that both electroweak theory and QCD are examples of USEFUL theories, and that more understanding (if not an overhaul of some basic principles) will be needed before a SUPERB theory of fundamentical particles is to emerge.

SOME MODERN TENTATIVE THEORIES

There are many attempts to find such a theory, but none of those can rise above the category TENTATIVE, since none have any significant experimental support (beyond any observational evidence that may have been part of the input into the theory in the first place). Among the most 'conventional' of these TENTATIVE theories are the so-called *grand unified* (or GUT) theories. These attempt to provide the same kind of unification between electroweak theory and QCD that electroweak theory itself achieved between the earlier theories of weak interactions with QED. In all these theories, it is assumed that the various physical interactions concerned are all part of the same scheme, interrelated by a large symmetry group. This symmetry would be exact at extremely high temperatures, but a process of 'symmetry breaking' sets in at lower temperatures which enables the different interactions to be distinguished from one another, and their smaller symmetry groups 'frozen out'. (The Higgs–Kibble procedure mentioned above makes use of this. Moreover, the TENTATIVE *inflation* theory of the early universe that I briefly mentioned in relation to the standard model of the big bang relies critically on this procedure.)

A TENTATIVE idea that was fashionable some ten or so years ago was a development of a much older theory (put forward by Kaluza and Klein in the 1920s) whereby space–time was envisaged as having more than the four dimensions that we are physically aware of, the extra dimensions being 'compactified', which here means that they are, in a sense, curled up into a tight ball of tiny size. The compactified dimensions are supposed to give rise to physical interactions, separate from the interaction of gravitation that arises from the ordinary four space–time dimensions in accordance with Einstein's general relativity. (The original Kaluza–Klein theory was five-dimensional, with one extra dimension to describe electromagnetism.) To get a picture of this kind of thing, imagine the surface of a hosepipe. On a large scale the pipe resembles a *one*-dimensional curve (to be thought of analogous to the four-dimensional space–time), but on closer examination, we find that the hosepipe surface is actually *two*-dimensional, one of the dimensions being compactified in this way. (One of the most popular Kaluza–Klein-type theories was eleven-dimensional, with seven dimensions compactified in the form of a 'squashed 7-sphere'.)

Another (still currently) popular idea is that of *supersymmetry*, according to which each particle of nature is accompanied by a sister particle whose spin differs from its own by half a unit. The idea of supersymmetry is based on the study of certain mathematical kinds of number that are intriguing (and, in some contexts, important) in their own right. When combined with the ideas of general relativity, supersymmetry gives rise to a theory called *supergravity*. One of the initial advantages of supergravity theory over standard general relativity was supposed to have been that it made the resulting quantized gravitational theory renormalizable (when it had not been before). However, these initial optimistic hopes have not been fulfilled and the original supergravity theory has somewhat fallen from favour.

Combining ingredients from all the above TENTATIVE theories is the (still) very fashionable (super)string theory. The basic *new* ingredient is a revival of an ingenious idea that was put forward by Nambu in 1970 in order to explain certain puzzling phenomena of strongly interacting particles. The theory was revived in a different form by Green and Schwarz in 1984, partly with an eye to describing gravitational interactions at the basic level, rather than the strong interactions, so as to provide a quantum gravity theory free of the non-renormalizable infinities that seemed to occur with standard quantization procedures applied to Einstein's theory. There have been strong claims, from time to time that not only does string theory provide a finite theory of quantum gravity, but that it also gives us a 'theory of everything'. Though I have no doubt that such claims are overstated and that the theory is firmly in the category TENTATIVE, having no support whatever on the observational side, it does have some points of considerable interest, and a brief description of the basic ideas is not at all out of place here.

The central idea is that the basic physical entities, or 'particles' in the theory are not the ordinary pointlike particles of conventional theory but extended objects that are to be thought of as little one-dimensional loops (or perhaps little open curves). In space–time terms, these so-called *strings* will describe two-dimensional surfaces, replacing the one-dimensional world-lines of conventional relativity theory. These surfaces are to be treated as the ingredients of a quantum field theory, and in Fig. 21 I have indicated the kind of way in which these surfaces would replace the lines of a standard Feynman diagram. One of the most intriguing features of these string surfaces is that several apparently different processes, that in Feynman diagram terms would have to be treated as different, may turn out to be actually the same from the point of view of string theory. (This was an important property of the original Nambu string theory, which explained certain observational features of strongly interacting particles, although in the modern versions of string theory, this connection with observational matters has become obscured.) An important aspect of string theory is that most of the infinities that arise in conventional quantum field theory now disappear. The frequent over-optimistic claims that string theory provides (in particular) a finite quantum

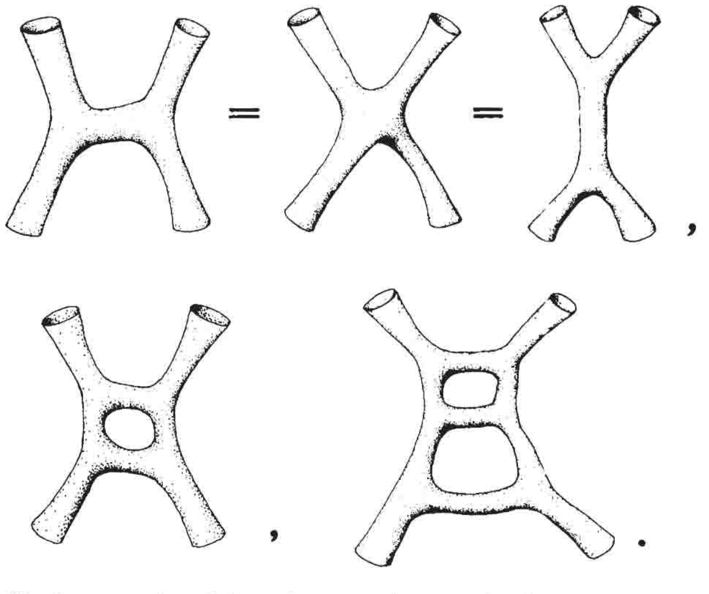


Fig. 21 Various string histories are depicted, the totality of which must occur in superposition. The upper three string histories all represent the same process since they can be deformed into one another.

gravity theory arose from this kind of reason. However, there are still infinities (and also conceptual problems) that arise when one tries to add together the effects due to string surfaces with all different possible topologies (which the theory tells us we are supposed to do). An even more serious conceptual problem is that the theory (at least in most of its original versions) required that the strings propagate through a space–time of twenty-six dimensions, or else of ten dimensions if the appropriate ideas from supersymmetry are also incorporated (to give *super*string theory). In fact, for various other reasons also, (super)string theory is very far indeed from providing an adequate quantum gravity theory, let alone a theory of everything.

On the other hand, there are important *mathematical* reasons why string theory has been taken seriously as a model that could perhaps provide some deep insights into the workings of nature at the sub-particle level. These reasons arise largely from the relation between the string surfaces and complex numbers. Recall the Riemann sphere that I referred to in connection with the spin of an electron, and with the way in which the complex numbers of quantum theory relate to the geometry of directions in physical space. The Riemann sphere is the simplest example of what may be called a *complex curve*, or complex one-dimensional space. In the complex sense, it is *one-*, rather than two-dimensional space because, as in the case of the Argand plane, a *single* complex number takes over the role of a *pair* of real numbers. More complicated complex curves can have more elaborate topological structures, such as that of a torus (surface of a ring) or of a sphere with several 'handles'. Such complex curves are referred to as *Riemann surfaces*. The power of string theory, in mathematical terms, lies in the close connection between string surfaces and these Riemann surfaces.

There is another body of ideas that represents an attempt to relate the theory of complex spaces (but now of higher dimension) to space–time structure and quantum mechanics. This is *twistor* theory, and it is the approach that I have been most involved with myself. Insofar as twistor theory is concerned with providing a new physical picture of nature, it must belong to the category TENTATIVE also. (It certainly provides no new physical predictions that could be put to observational test.) But twistors may also be viewed in a more modest light than this, namely as giving a mere *mathematical* reformulation of *existing* physical theory, particularly in the context of special (or general) relativity. It provides unusual insights into Maxwellian electrodynamics and quantum field theory.

The two most important motivational ideas that are incorporated into twistor theory are, first, a fundamental use of *complex numbers* and complex spaces (as just referred to in connection with string theory) and, second, that there should be an essentially *non-local* description of space–time at the basic level (as would seem to have importance in relation to the space–time non-locality involved in the 'EPR paradox' referred to above). Twistors may, in the first instance, be viewed as entire light rays (photon histories) in space-time. If we take such a light ray to be represented as a single point in another space—which for the moment let us call *twistor space*, a space which we shall try to regard as more primitive than the space–time itself—then, in twistor terms, we can represent a space–time point conversely as the family of light rays that pass through that point, i.e. as a locus in twistor space representing such a family. This locus turns out to be naturally describable as a Riemann sphere, see Fig. 22. (The light rays through a particular space–time event describe the *celestial sphere* as seen by an observer at that event, and it turns out that the Lorentz transformations that one must apply on passing from one observer to another at the same event are identical with those transformations which preserve the complex-curve structure of a Riemann sphere.) The idea is to regard the whole of twistor space as a complex space. For Minkowski's space–time of special relativity, this works remarkably well—although to get the whole of twistor space, we must demand that the light rays represent the space–time histories not of simply massless point particles, but of extended objects that describe the spin and energy of a massless particle like a photon. (An actual photon does indeed spin left-handed or right-handed about its direction of motion.) This is not the place to describe the many subtleties and unexpected mathematical interrelationships that arise in twistor theory. The theory has turned out to have a number of different applications within areas of pure mathematics

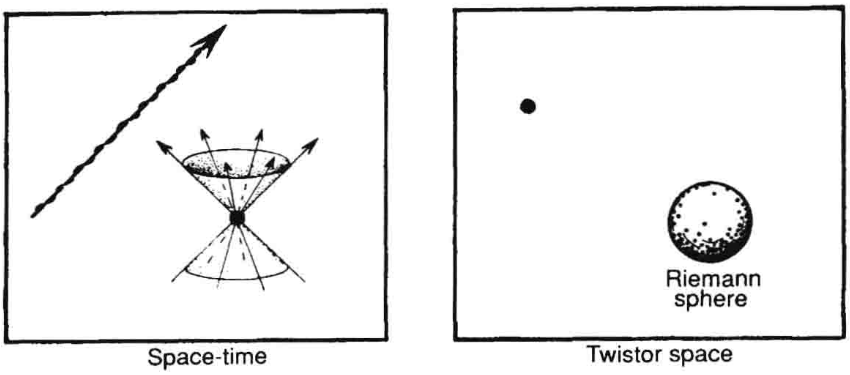


Fig. 22 The twistor correspondence: the space–time history of a (spinning) massless particle (*left*) is represented as a single point in twistor space (*right*); a point in space–time, described in terms of the light rays through it (*left*), is represented as a Riemann sphere in twistor space (*right*).

(mainly differential geometry, representation theory, and integrable systems). The main successes on the physical side have been in general relativity (although a full description of Einstein's equations remains fundamentally elusive) and in quantum field theory. Most significantly, Hodges has found a way of eliminating infinities in QED and electroweak theory, within the context of twistor diagram theory (the twistor analogue of Feynman diagram theory). So far, twistor theory provides only a body of techniques and it is not an independent physical theory. (A scheme due to Singer and Hodges, possibly leading to such a theory, arises if twistor ideas are combined with the ideas of *conformal field theory*—a mathematical theory using the Riemann surfaces from string theory—in higher dimensional form.)

There is no shortage of alternative TENTATIVE schemes (such as other higher-dimensional generalizations of string theory called *membrane* theory) and there is no consensus as to the form that the next SUPERB theory is likely to take, though most theoretical physicists are in agreement that there are indeed further SUPERB theories to be found. The combination of quantum theory with general relativity—the finding of the correct *quantum gravity* theory—is accepted as being one of the profound challenges of such a future theory, although not many physicists yet seem to share my own view that this will entail a fundamental change in the very structure of quantum theory in which **U** and **R** will be united in a single non-linear procedure. If one point of agreement can be found, however, it is the faith—implicit or explicit—underlying all these endeavours: that future successes must come through yet deeper investigations of the interrelations between mathematics and physics. Perhaps *The Modern Physicist's View of Nature* can be summed up as this faith: that there *is* indeed a profound harmony between mathematics and the workings of the natural world!

FURTHER READING

Some of this article is based on ideas described in my semi-popular book (Penrose), where further discussion may be found of some of the SUPERB theories described above. For a very thorough account of the development of physics over the past hundred years, see Pais. For popular accounts of the more recent (including USEFUL and TENTATIVE) theories, see Barrow, Close, Peat, and Squires.

J. D. Barrow, *The world within the world* (Oxford University Press 1988).

F. Close, *The cosmic onion: quarks and the nature of the universe* (Heinemann, London 1983).

A. Pais, *Inward bound* (Oxford University Press, Oxford 1986).

F. D. Peat, *Superstrings and the search for the theory of everything* (Contemporary Books, Chicago, IL 1988).

R. Penrose, *The emperor's new mind: concerning computers, minds, and the laws of physics* (Oxford University Press 1989).

E. Squires *To acknowledge the wonder* (Adam Hilger, Bristol 1985).

218

GRAVITY AND QUANTUM MECHANICS

General relativity and gravitation 13. Part 1: Plenary Lectures 1992. Proc. 13th Int. Conf. General relativity and gravitation (1992), Eds R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi, 179–89. Institute of Physics, Bristol, 1993

This is the write-up of my Plenary talk at GR13 (13th International Conference on General Relativity and Gravitation, Cordoba, Argentina, 1992), where I had managed to do a calculation the day before my talk (I believe it was) which served to alter my viewpoint with regard to the spontaneous reduction of a quantum state, due to effects of Einstein's general relativity; see my introductory comments to Chapter 215. Previously (see, for example Chapter 166, pp. [141]–[144]), I had been thinking in terms of a possible notion of 'gravitational entropy' or 'longitudinal graviton number' as what might signal a spontaneous quantum state reduction. Even as late as in Chapter 201 p. [614], I was arguing for a criterion in which a scale of mass (ultimately dependent on the Planck mass) at which quantum superpositions would be forbidden. The revised idea (of a general kind already put forward some years earlier by Diósi) was that there would be a *time scale* for collapse, which can be calculated from the amount of mass displacement between the states.

Gravity and quantum mechanics. *General relativity and gravitation 13. Part 1: Plenary Lectures 1992. Proc. 13th Int. Conf. General relativity and gravitation* (1992). Eds R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi, 179–89. 1993

Gravity and quantum mechanics

Roger Penrose
Mathematical Institute
Oxford UK

0 Introduction

At a conference in honour of John S. Bell, held in October 1991, the following quotation, from a paper by Bell and Nauenberg (1966), was presented (in a talk by Dipankar Home):

> "... the quantum mechanical description will be superseded. In this it is like all theories made by man. But to an unusual extent its ultimate fate is apparent in its internal structure. It carries in itself the seeds of its own destruction."

I was struck by a similarity of sentiment, as expressed here, with one that I have myself often expressed, but now in relation to general relativity; e.g. (Penrose 1991):

> "... in a clear sense, general relativity predicts its own downfall as a complete description of the structure of space-time"

There is, indeed, a remarkable parallel, in this regard, between these two great physical theories. Both theories are now known to be exceptionally accurate, within the range of phenomena to which they are applied; yet both present us with profound difficulties. In the case of general relativity, the profound problem is that of *space-time singularities*, whose presence in Einstein's theory is an implication of the theory itself (cf. e.g. Hawking and Penrose 1970). In the case of quantum theory, the difficulty is the so-called *measurement problem*, which still has no really satisfactory solution. Many different viewpoints are expressed in relation to the measurement problem, often accompanied by claims of some kind of solution - but where the "solution" proposed satisfies none of those holding to an opposing viewpoint.

Despite the fact that *both* theories have their profound difficulties, the normal attitude to them, amongst contemporary physicists, is very different in the two cases. The standard reply to the question of the space-time singularities of classical general relativity is that that theory should be modified by applying to it, in some appropriate way, the rules of standard quantum theory - or, if this fails to work, then the classical Einstein theory itself should be changed (as would be the case with Kaluza-Klein-type theories, supergravity, and the classical limit of superstring theory) so as to force it into a more amenable shape for the consistent application of quantum procedures. When it comes to quantum theory's own problems, on the other hand, the normal attitude among contemporary physicists seems to be that the problem ought just to go away - if we only understood the theory itself properly. There is little suggestion that the very rules of quantum mechanics might in any way be in need of modification, or that quantum mechanics might itself derive any benefit from general relativity. (Isolated expressions of dissent from this general viewpoint have, however, been put forward from time to time; cf. Károlyházy 1966, 1974, Károlyházy, Frenkel, and Lukács 1986, Komar 1969, Diosi 1989, Penrose 1981, 1989.)

My own attitude to these problems is a much more even-handed one than that normally adopted. I believe that the sought-for union between general relativity and quantum theory will involve as much change in the structure of quantum theory as in general relativity. I believe that *both* quantum theory and general relativity are (superb) approximations to some hitherto undiscovered new theory, where each would be valid in some appropriate limiting sense. The solution to the singularity problem and to the measurement problem would then both find their resolutions within this new theory; indeed, the solutions to these two problems should arise, accordingly, as two sides of but a single coin.

1 Quantum theory's fundamental problem

Since the singularity problem of general relativity is now generally accepted as providing a limit to that theory's classical applicability, I shall concentrate here on the central problem of quantum theory. I think that a very profound remark concerning different people's attitudes to quantum theory was made to me some years ago in a dinner-table comment by Bob Wald:

> "If you really believe in quantum mechanics, then you can't take it seriously."

This expresses the fact that it is only those who dissent from the standard "Copenhagen" view who are prepared to regard the state-vector $|\psi>$ as actually representing (even an approximation to) physical *reality*. Niels Bohr, on the other hand, was one of the strongest proponents of the idea that $|\psi>$ was to be regarded merely as a calculational tool, an expression only of our knowledge about a physical system, and to be used simply for the mathematical calculation of probabilities with regard to the various results of "measurements" that might be performed on a system. Bohr, indeed, was someone who really did believe in quantum mechanics, and so was unable to take $|\psi>$ seriously as a description of actual physical reality.

Those who *do* take $|\psi>$ seriously as an objective description of the physical world - although possibly only a provisional or approximate one, perhaps to be superseded if quantum theory is someday replaced by an even more accurate theory - seem to me to fall into two camps. There are those who believe that the present theory of the way that $|\psi>$ evolves, namely *unitary evolution* **U**, must be preserved at all costs, and that the phenomenon of *state-vector reduction* **R** is some kind of illusion; on the other hand there are those who believe that **U** applies only to an approximate (though highly accurate degree) and that **R** represents a real physical phenomenon that effectively interrupts, from time to time, the continuous evolution according to **U** (such as might be entailed by some modification of the Schrödinger equation). In my own view, it is those in this latter camp (and I would count myself among their number) who are taking the formalism of quantum mechanics most "seriously" of all, because they believe that *both* the main ingredients of the theory, indeed **R** as well as **U**, are to be taken seriously as describing something objectively real about the evolution of the world. Among those in the latter camp would be such as Károlyházy 1974, 1986 and his co-workers, Pearle 1985, 1989, Ghirardi, Rimini, and Weber 1986, Diosi 1989 moreover, I would count John Bell as essentially being in this camp.

Belonging to the former camp within the group who "take $| \psi >$ seriously" would be those who follow a "many-worlds" viewpoint with regard to the quantum state. Accordingly, $| \psi >$ always evolves according to **U**, but all the different outcomes of a "measurement" must co-exist in different "worlds", each being perceived by a separate copy of any observer. The difficulties with this viewpoint lie, to my own mind, not so much in its lack of economy and in its extreme stretching of one's notions as to what "reality" should encompass, but in its incompleteness with regard to its descriptions as to what a conscious observer would actually perceive about the world and about the probability values that such an observer would assign to different perceived events. In short, in itself, it provides no solution to the "Hilbert space basis problem" or to the problem of why squared moduli of amplitudes should actually become probabilities. Also in this former camp might be those who hold to some form of "decoherence" explanation of R, although the standard explanations of this kind, described by John Bell as FAPP ("for all practical purposes") explanations, could really be satisfying only to those who do *not* take $| \psi >$ seriously as providing an actual description of the physical world. With regard to those "decoherence" viewpoints, such as those put forward by Griffiths (1984), Gell-Mann, and Hartle (1990), which adopt a path-integral-type picture of reality, I would regard them as being to some extent *modifications* of standard quantum theory (which would place them in the latter camp), but in any case, as not providing a real resolution of the measurement problem. For they accept their own versions of **U** and **R** as things with distinct mathematical descriptions, and they do not tell us under what actual physical circumstances a "measurement" would be deemed as taking place.

2 Two sides to the state-reduction phenomenon

Consider a simple (thought) experiment, where a photon source (the lamp L) and a detector (the photo-cell P) are placed at opposite ends of a hall, with suitable paraboloidal or ellipsoidal mirrors placed so as to ensure that virtually every photon emitted by the source would reach the detector, provided that there is no obstruction on the line joining them (fig. 1). Let us now imagine that there is a half-silvered mirror M placed mid-way between them on this line, tilted at 45° to the line. We are to take it that any photon moving along this line in *either* direction, when reflected off the mirror, would be absorbed at a point of the hall wall (at B if the photon comes from the direction of L, and at A if the photon were to come from P). Suppose that from time to time photons are emitted by L and that from time to time P (assumed to be a 100% efficient detector) registers the reception of a photon. Assume, also, that every time L emits a photon, it registers this fact (again with 100% efficiency), so that all the emission and reception events are clear-cut *measurements*.

Suppose, now, that it is given that L has emitted a photon. We ask: what is the probability that P receives one? Clearly the answer is $\frac{1}{2}$. This is a consequence of the standard quantum rule whereby probabilities are obtained by squaring the moduli of complex amplitudes. The photon wave function splits into two components, each of which has an amplitude $\frac{1}{\sqrt{2}}$ (times a possible phase factor). One component reaches the detector at P, and the other reaches the wall at B. The squared modulus of each is $\frac{1}{2}$, so the respective probabilities of reaching P or B are each $\frac{1}{2}$.

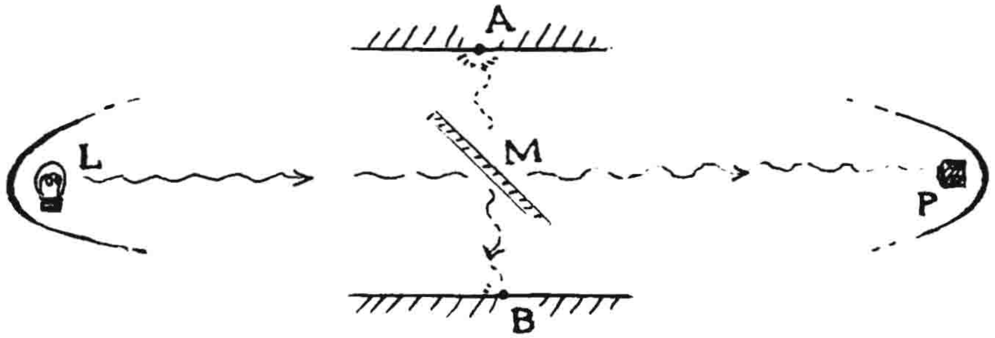


Fig.1 Photons, aimed towards the detector P, are emitted from time to time by the source L. Between the two is a half-silvered mirror M which partially deflects the photons to the absorbing wall at B. A photon ejected from the wall at A also could reach P. The probability that P receives a photon, given that L emits one, is governed by the quantum-mechanical squared-modulus rule, whereas the probability that L has emitted one given that P receives one is determined by the second law of thermodynamics.

Let us now ask a time-reversed kind of question. Suppose that we are given that P has received a photon. We now ask: what is the probability that L had emitted one? Now the answer is certainly not $\frac{1}{2}$, but we have a probability of essentially 1 that the photon came from L and a probability of essentially 0 that it came from A, which is the only other possibility. It is virtually certain that the photon came from the lamp L rather than from the wall at A. However, had we tried to use the "squared modulus" rule for calculating these probabilities, we should have indeed obtained $\frac{1}{2}$ for each of them. This discrepancy has nothing to do with the fact that it is usual to evolve wave functions into the future rather than into the past. Precisely the same answers are obtained whichever way we evolve. The conclusion is that the "squared modulus" rule simply does not work if we try to apply it to obtain retrospective probabilities. There is no real reason why it should. The miracle is that such a simple and elegant rule indeed works in the future time-direction!

If we were to ask what rules indeed govern probabilities in the past time direction, then we are forced to consider such matters as the second law of thermodynamics. This law is certainly playing a role here, because for a photon to jump out of the wall at A in order to be reflected off the mirror and arrive at the detector P, a severe violation of the second law would be needed. Basically it is the second law that is responsible for a virtual vanishing of the probability of a photon emerging from the wall at A.

It is clear from all this that there is no necessity for the probabilities in the past and future directions to match up in a particular experimental situation, such as this. However, there are strong reasons for expecting that if we were to consider, in some appropriate sense, the totality of *all* possible "experimental situations", then there would indeed be a past-future balance for the probabilities, taken as a whole. Basically, we must balance the physics responsible for the second law of thermodynamics against the physics responsible for the quantum-mechanical probabilities - in order to show that these two areas of physics must actually be two different aspects of the *same* physics.

Gravity and quantum mechanics. *General relativity and gravitation 13. Part 1: Plenary Lectures 1992. Proc. 13th Int. Conf. General relativity and gravitation* (1992). Eds R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi, 179–89. 1993

3 The role of the Weyl curvature hypothesis

I do not wish to repeat the entire argument here, since it is one that I have given many times before (see, particularly, Penrose 1981, 1989). The essential point is that the second law arises ultimately from the enormous constraint on the space-time geometry that was operative at the big bang singularity. This initial constraint (which for a 10^{80}-baryon universe would amount to a restriction of the phase-space down to a region whose volume is about $\exp(-10^{123})$ of that of the entire phase space) starts the universe off with a very tiny entropy, as compared with what it "might have been", and the entropy has been rising ever since, in accordance with the second law. The simplest way to impose such a constraint is to take it as a geometrical restriction: the initial Weyl curvature is to be zero - or at least to be much smaller, in some appropriate sense, than it would have been for a generic big bang. This restriction, which is taken to apply only to *initial*, and not to final space-time singularities (so that we can derive the time-asymmetric second law), is referred to as the *Weyl curvature hypothesis* (WCH). (In recent work, R.P.A.C.Newman has shown that a form of WCH due to K.P.Tod can be used to derive an initial Friedmann-Robertson-Walker structure for the early universe, assuming an appropriate perfect-fluid state.)

Thus it is WCH (or something closely of this nature) that we seem to have to balance against the "squared modulus" rule of the **R** part of quantum mechanics. Moreover, if we take it that this singularity structure, as implied by WCH, is a feature of the correct quantum gravity theory - or rather, of the putative correct *new* theory out of which classical general relativity and standard quantum mechanics must both emerge as appropriate limiting cases - then we appear to conclude that it is this new theory that must also be responsible for the probabilities involved in **R**. In other words, the modification of quantum theory that would be needed in order for us to be able to understand **R** as a real physical process must be a modification that operates at the level where gravitational effects begin to become quantum-mechanically important.

In fact the connection can be made more explicit if we consider the "Hawking box" thought experiment (described in Penrose 1981, 1989). In this situation, the phase-space volume gain that occurs with experiments like that discussed in the previous section, where in effect there are more possible outputs than there are inputs (here: P and B allowed as outputs, whereas L is allowed as an imput but A is effectively forbidden), is balanced against a phase-space loss that occurs in black holes. There is a net loss because black holes are allowed by WCH, but their time-reverses - the white holes - are forbidden by it. The black hole singularity is a future singularity which absorbs information, whereas a white hole's singularity would be a past singularity, like the big bang, but with an infinite rather than zero Weyl curvature. The absorbing of information by a black hole's singularity is what is responsible for the Hawking effect, according to Hawking's original derivation (fig. 2). It should be mentioned, however, that there are some opposing points of view, according to which it is argued that if the entire history of a black hole's formation and ultimate disappearance (or possible non-disappearance) due to its Hawking evaporation is taken into account, then information is not actually lost. My own considered opinion is that such information restoration is not plausible, particularly if one is to believe that something of the nature of WCH must hold, so that white-hole singularities, with their potential for creating new phase-space volume, are not permitted.

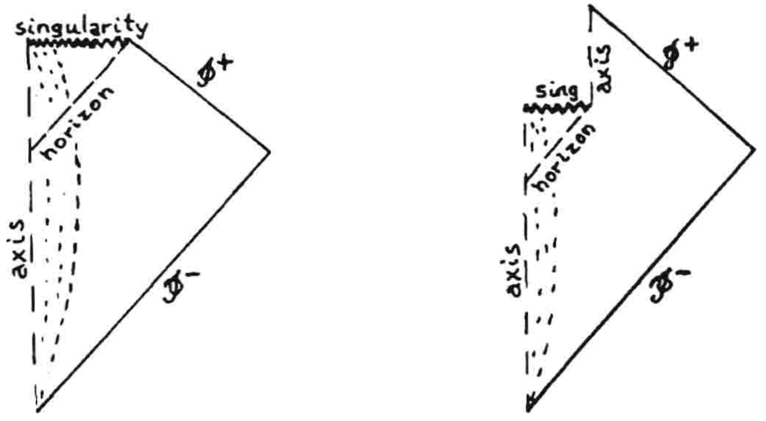


Fig.2 Hawking evaporation. The conformal diagram on the left gives the geometrical background for Hawking's original derivation of the presence of Hawking radiation - as an effect of a loss of information in the Black hole's singularity. The diagram on the right takes into account the final disappearance of the hole from the back-reaction of this radiation on the space-time geometry.

4 A gravitational origin for R?

An implication of the preceding discussion is that we should look for some criterion for the onset of quantum state-vector reduction which is of a gravitational character. There are, indeed, other reasons for suspecting that the standard quantum formalism might not apply without change in situations where the curved-space features of general relativity begin to become significant. For example, the normal ideas of energy, momentum, and angular momentum, in a quantum context, relate these physical concepts directly to symmetries of the space-time manifold, whereas in a general-relativistic setting, such symmetries would normally be absent. This leads to certain severe difficulties with regard to quantization, since it is indeed the quantum rules for energy, momentum, and angular momentum that provide the initial guide to quantization in a normal flat-space setting. (We recall that the passage from a classical to a quantum description - i.e. from a symplectic manifold to Hilbert space - is not a well-defined mathematical procedure, except in the presence of further structure such as that supplied by space-time symmetries; cf. Woodhouse 1980.) Even the notion of positive/negative energy (i.e. frequency) splitting, vital for the setting up of quantum field theory, is not well-defined in a general curved space-time.

These difficulties occur even for the problem of quantizing within a curved space-time background. As is well known, the problems that arise when it is the curvature of space-time itself that has to be subjected to the laws of quantum mechanics are of a much more serious nature. One must ask, in particular, how it is possible to interpret a physical system in which different space-time geometries are being subjected to quantum linear superposition; indeed, even worse than that, one must ask how is it possible to interpret *other* physical objects that try to inhabit such a curious quantum-superposed background? When there is no natural correspondence even between the individual points in the different classical space-times that are to be superposed, then it is hard to see how interference between the different physical states associated with the different geometries can be understood.

What has this to do with state-vector reduction **R**? Let us consider a type of situation like a "Schrödinger's cat", in which one strives to produce a state in which a pair of macroscopically distinguishable alternatives are linearly superposed. For example, in Fig. 3 we have a situation in which a photon impinges upon a half-silvered mirror, and the photon state becomes a linear superposition of being transmitted through it and reflected by it. The transmitted part of the photon's wave function activates (or would activate) a device which moves a macroscopic spherical lump from one location to another. So long as Schrödinger evolution **U** holds good, the "location" of the lump becomes a quantum superposition of its being in the original position with its being in the displaced position. As soon as **R** comes into effect, we are allowed to consider that the lump is in either one position or the other - and a "measurement" has been performed. The idea here is that this is an entirely objective physical process which occurs whenever the mass of the lump is large enough or the distance it moves is far enough. In particular, it has nothing to do with whether or not a conscious observer may happen to have actually "observed" the movement or otherwise of the lump. (In this, I am imagining that the device that detects the photon and moves the lump is itself "small" enough that it can be treated entirely quantum mechanically, and it is only the lump that registers the measurement. For example, in an extreme case, we might simply imagine that the lump is poised sufficiently unstably that the mere impact of the photon would be adequate to cause it to move off significantly.)

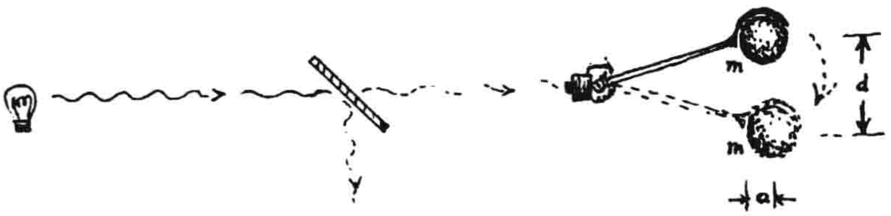


Fig.3 A photon impinges on a half-silvered mirror, so that only the transmitted part of its wave function is received by a device which if activated would move a macroscopic spherical lump of mass m and radius a through a distance d. Is there stage at which linear superposition between the two possible positions of the lump fails, and the lump actually becomes localized in one position or the other?

How would such a situation be treated according to the standard **U** procedure of quantum mechanics? After the photon has encountered the mirror its state would have to be considered as a non-local system, with the two parts of its wave function in two very different locations. One of these parts then becomes entangled with the device and finally with the lump. We thus have a quantum state which involves a linear superposition of two quite dfferent positions for the lump. Now the lump will have its gravitational field, which must also be involved in this superposition. Thus, the state also involves a superposition of two different gravitational fields - i.e., according to Einstein's theory, with two different space-time geometries! The question is: is there a point at which the two geometries become sufficiently different from each other that the rules of quantum mechanics must change, and rather than forcing the different geometries into superposition, Nature chooses between one or the other of them to effect the reduction procedure **R**?

5 The weak-field gravitational symplectic integral

As a guide to establishing a plausibe criterion, in accordance with the foregoing general ideas, we consider a certain integral (Fierz 1940), over a spacelike Cauchy hypersurface Σ, which determines the symplectic structure on the function-space of weak vacuum gravitational fields (i.e. massless fields of spin 2 in flat space). This integral (up to a real factor) is:

$$\{K_{(1)}, K_{(2)}\} = \int_\Sigma (\Psi_{(1)ABCD}\bar{\eta}_{(2)A'}^{BCD} + \bar{\Psi}_{(1)A'B'C'D'}\eta_{(2)A}^{B'C'D'})d^3x^{AA'},$$

where the fields (labeled by (1) and (2), respectively) are described by linearized curvatures in Minkowski space

$$K_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{A'B'C'D'},$$

where we have a (Dirac) chain of potentials

$$\begin{aligned} \nabla_{BB'}\eta_A^{B'C'D'} &= \chi_{AB}^{C'D'}, & \nabla^{AA'}\eta_A^{B'C'D'} &= 0,\\ \nabla_{BB'}\chi_{AB}^{C'D'} &= \gamma_{ABC}^{D'}, & \nabla^{AA'}\chi_{AB}^{C'D'} &= 0,\\ \nabla_{BB'}\gamma_{ABC}^{D'} &= \Psi_{ABCD}, & \nabla^{AA'}\gamma_{ABC}^{D'} &= 0,\\ & & \nabla^{AA'}\Psi_{ABCD} &= 0, \end{aligned}$$

all of Ψ_{ABCD}, $\gamma_{ABC}^{D'}$, $\chi_{AB}^{C'D'}$, $\eta_A^{B'C'D'}$ being symmetric in their unprimed and primed indices separately. (See Penrose and Rindler 1984 for the relevant spinor notation.) In vacuum, the divergence of the integrand vanishes, showing that (with suitable fall-off at infinity), the integral is independent of the choice of Σ. The symplectic form $\{,\}$ is closely related to the scalar product $<|>$. We have $< K \mid K >= i\{K, \mathbf{J}K\}$, where **J** multiplies the positive-frequency part of K by i and the negative frequency part by -i. Integrating by parts, we find, schematically, (for fields falling off suitably at infinity):

$$\int \Psi_{(1)}\bar{\eta}_{(2)} = -\int \gamma_{(1)}\bar{\chi}_{(2)} = \int \chi_{(1)}\bar{\gamma}_{(2)} = -\int \eta_{(1)}\bar{\Psi}_{(2)},$$

whence

$$\{K_{(1)}, K_{(2)}\} = -\{K_{(2)}, K_{(1)}\}.$$

The quantity γ... describes the linearized spin-coefficients or, equivalently, the linearized Christoffel symbols; moreover, the Hermitian part of χ... describes the linearized metric.

We can use this symplectic integral as a measure of the *difference* between two weak (vacuum) gravitational fields. When this difference reaches order unity, all quantities being measured in *absolute units* ($G = c = \hbar = 1$), we can say that the two fields are "significantly different". The idea is, roughly, that when two (weak) gravitational fields are judged as being significantly different according to this criterion, then quantum linear superposition between them cannot be maintained, and the state reduces to one or the other of them - so **R** has been effected!

6 A gravitational criterion for the onset of R

How can we apply this kind of idea to the situation considered in **4.**, as illustrated in Fig. 3? A difficulty is that the symplectic integral of **5.** applies (and is divergence-free) only in the vacuum region, whereas we need it in situations where the lumps are actually present. Any spacelike Cauchy hypersurface Σ must intersect the lump (in both its alternative locations). One point to note, however, is that there is actually no need for Σ' to be spacelike everywhere. It just has to be topologically deformable to a Cauchy hypersurface. As it stands, this is no direct help, but it indicates that we need not feel restricted to spacelike hypersurfaces in applying the integral expression.

What seems suggestive, however, is to perform the integral over a *timelike* hypersurface Σ' in a region between the two locations of the lump after they have separated from one another, as indicated in Fig. 4. The hypersurface Σ' is bounded between two times, and let us call the time-interval between them T. A tentative suggestion might now be that the system reduces in a timescale T such that the symplectic integral, applied between the gravitational fields of the two possible positions of the lump, is of order unity.

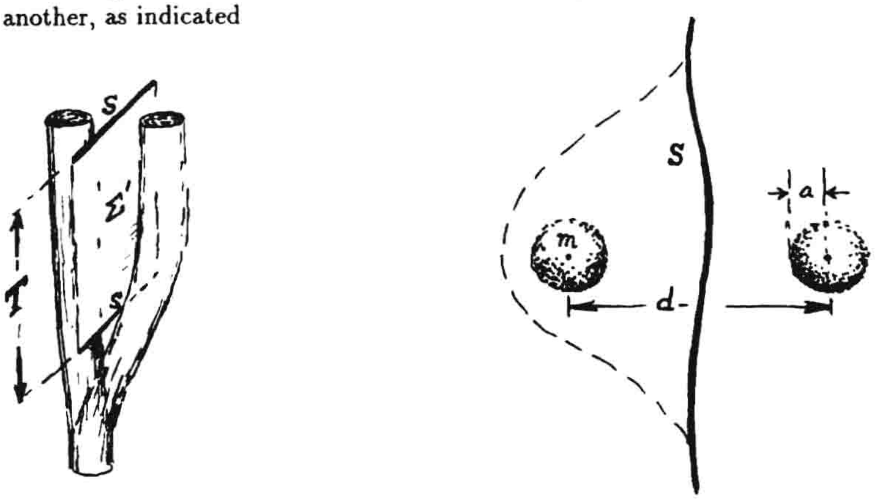


Fig.4 The two alternative locations of the lump, as depicted in Fig.3, are indicated, on the left, in a space-time diagram. The symplectic integral is performed over a portion of timelike hypersurface, suggesting that a criterion for reduction to take place is when this integral becomes of order unity. The spatial configuration is depicted on the right.

The situation is basically a Newtonian one, assuming that, compared with the speed of light, the lump moves slowly, and the gravitational escape velocity at its surface is also tiny. In the Newtonian limit, for an essentially static situation, and taking a hypersurface Σ' that is also static - i.e. the product of a spacelike 2- surface S with an interval of the time axis (of duration T) - our symplectic integral becomes (some simple constant multiple of)

$$T \int_S (\phi_{(2)}\vec{\nabla}\phi_{(1)} - \phi_{(1)}\vec{\nabla}\phi_{(2)}) \bullet d\vec{x}.$$

Here ϕ denotes the Newtonian gravitational potential due to the lump. The parenthetic suffix specifies the lump location. We note that the integrand is divergence-free in vacuum, so the integral remains unchanged under topological deformations of S through regions of matter-free space. Where lump matter is present (density ρ), we make use of

$$\vec{\nabla} \bullet (\phi_{(2)}\vec{\nabla}\phi_{(1)} - \phi_{(1)}\vec{\nabla}\phi_{(2)}) = -4\pi\phi_{(2)}\rho_{(1)} + 4\pi\phi_{(1)}\rho_{(2)};$$

thus, moving S across one of the two possible locations for the lump, we find that the value of our integral is $4\pi T$ times the gravitational energy of one instance of the lump in the gravitational field of the other. This would suggest that the time that it takes for the state to reduce is a simple multiple of the reciprocal of this energy.

This energy is

$$m^2/d,$$

where the lump has mass m, and the spatial displacement between the centres of its two alternative positions is d. However, the reciprocal of such a measure is not really a satisfactory suggestion as to the time it takes for reduction to occur in this situation, because for large displacements the measure gets smaller and tends to zero. Thus the suggested "time" would get longer, making it *less* likely for **R** to occur, the greater the displacement, rather than *more* likely as one would expect. Moreover, when the two instances of the lump overlap, there would be no way of locating Σ' so that it lies entirely in the vacuum region.

This suggests that the reduction time should be something a little different from this, but perhaps belonging to the same order of ideas. In fact a slight modification of this expression (arrived at some ten weeks following the Cordoba meeting) does give something reasonable. The expression m^2/d is the gravitational energy gained (in absolute units) if one moves one instance of the lump in from infinity, in the gravitational field of the other, until it reaches the required separation d. If we consider, instead, the more relevant energy of moving one instance of the lump away from the other, starting from coincidence, until they reach the required separation, then, taking the lump to have radius a, we obtain something of the order of m^2/a for this energy. Although the energy now indeed increases as the separation increases, the additional energy in moving from the contact position to all the way out to infinity is of the same order as that in moving from coincidence to the contact position. Thus, as far as orders of magnitude are concerned, one can ignore the contribution due to the displacement d, and take the reduction time to be of the order of

$$a/m^2.$$

It is reassuring that this gives very "reasonable" answers in certain simple situations. For example, in the case of a nucleon, where we take a to be 10^{-13}cm, which in absolute units is about 10^{20}, and m to be about 10^{-20}, we get a reduction time of 10^{60}, which is about a Hubble time. If we consider a droplet of water of radius 10^{-5}cm, we get a reduction time of about a day; if of radius 10^{-4}cm, the reduction time, according to this scheme is roughly a second; if of radius 10^{-3}cm, then about 10^{-5} of a second. So far, this seems to be quite plausible, but clearly more work is needed to see whether the idea will survive more stringent examination.

I am grateful to many colleagues for valuable comments, most particularly Abhay Ashtekar and Ted Newman.

322 Gravity and quantum mechanics. *General relativity and gravitation 13. Part 1: Plenary Lectures 1992. Proc. 13th Int. Conf. General relativity and gravitation* (1992). Eds R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi, 179–89. 1993

References

Bell, J.S. and Nauenberg, M. (1966) The moral aspect of quantum mechanics. In *Preludes in Theoretical Physics* (eds. A. De Shalit, H. Feshbach, and L. Van Hove; North-Holland, Amsterdam; pp.279- 86). Reprinted in: Bell, J. S. (1987) *Speakable and Unspeakable in Quantum Mechanics.* (Cambridge Univ. Press, Cambridge).

Diosi, L. (1989) *Phys. Rev.* A40, 1165.

Fierz, M. (1940). Uber den Drehimpuls von Teichlen mit Ruhemasse null und beliebigem Spin. *Helv. Phys. Acta* 13, 45-60.

Gell-Mann, M. and Hartle, J.B. (1990) in *Complexity, Entropy, and the Physics of Information, SFI Studies in the Science of Complexity, Vol. VIII* (ed. W. Zurek; Addison Wesley, Reading).

Ghirardi, G.C., Rimini, A., and Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Phys. Rev.* D34, 470

Griffiths, R. (1984) *J. Stat. Phys* 36, 219.

Hawking, S.W. and Penrose, R. (1970) The singularities of gravitational collapse and cosmology, *Proc. Roy. Soc. (London)* A314, 529-548.

Károlyházy, F. (1966) *Nuovo Cim.* A42, 390.

Károlyházy, F. (1974) Gravitation and quantum mechanics of macroscopic bodies, *Magyar Fizikai Polyoirat* 12, 24.

Károlyházy, F., Frenkel, A. and Lukács, B. (1986) On the possible role of gravity on the reduction of the wave function in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Komar, A.B. (1969) Qualitative features of quantized gravitation *Int. J. Theor. Phys.* 2, 157-60.

Pearle, P. (1985) 'Models for reduction', in *Quantum Concepts in Space and Time*, eds., C.J. Isham and R. Penrose, (Oxford Univ. Press, Oxford).

Pearle, P. (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. *Phys. Rev.* A39, 2277- 89.

Penrose, R. (1981) Time-asymmetry and quantum gravity in *Quantum Gravity 2*, eds D.W.Sciama, R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Penrose, R. (1989) *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, (Oxford Univ. Press, Oxford).

Penrose, R (1991) What does the big bang tell us about quantum gravity? *Mem. S.A.It.* 62, No.3, pp.607-614.

Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time*, Vol. 1: Two-Spinor Calculus and Relativistic Fields (Cambridge University Press, Cambridge).

Woodhouse, N.M.J. (1980) *Geometric Quantization* (Clarendon Press, Oxford).

219

FOREWORD TO *Mr TOMPKINS IN PAPERBACK*

George Gamow (Canto edition). Cambridge University Press, 1993

As a rule, forewords that I have written have not been included in these Collected Works, but exceptions have been made when such a foreword expresses a scientific point that the editors of the Collected Works have considered to be of interest. I am glad that this rule has been applied positively in this case, because when I was at the age of around ten, or so, my older brother Oliver acquainted me the Mr Tompkins books, reading me large sections. I found them fascinating and compelling, and they had an undoubted influence in later steering me to work in the physical sciences. I hope that this foreword repays a little of my debt to George Gamow.

Foreword

It is an especial pleasure to me that George Gamow's two gems—concerning the two adventures of Mr Tompkins in the wonderful worlds conjured up when the speed of light is made small or Planck's constant large—are now reprinted in paperback. Whilst, with hindsight, one may find details to quibble about, the excitement of these deeply delightful tales remains as fresh with me as it was some fifty years ago when I first encountered them. Though physics has moved on in many ways, the basic physics of relativity and quantum theory has not changed. By his ingenuity and narrative skills, Gamow is able to transform some of the puzzling and obscure mysteries of this basic physics—a physics which, indeed, is still modern—into magical and enthralling stories for children.

I remember reading (or being read) the Tompkins stories as a quite young child, and I am sure that their magic was responsible, to a very considerable extent, for the great excitement that fundamental physics has held for me for the rest of my life. I still vividly recall the tigers of the quantum jungle, and the old woodcarver's boxes of mysterious coloured balls (the nucleons), the relativistically flattened bicycle, and the professor calling out 'Just lie down and observe' as he and Mr Tompkins see their miniature universe collapse inwards upon them. It was Mr Tompkins who made the new physics vivid and real for me as a child and I am sure that he will continue to do the same for a great many others.

Written in the early 1940s (updated in 1965), Gamow's main concerns were the general principles of relativity and cosmology, of quantum theory and particle physics. Has his account become outdated? Except in certain details, the essentials of this physics have not changed, and so Gamow's descriptions remain, to this day, essentially modern. Perhaps the main ways in which physics has moved forward since Gamow's time relate to particle physics. More particles are known now than in Gamow's day, but there are now good theories to describe them. We know something about neucleon substructure (in terms of quarks) and the strong and weak interactions of particles (in terms of the 'gauge theories' that underlie what is now known as the 'standard model'). The neutrino, whose existence was still enigmatic as Gamow first wrote, is now well established both theoretically and experimentally. With regard to relativity, Gamow's descriptions of flattened bicycles and city blocks are intuitively helpful, but they do not represent what an observer would actually *see*. Taking the finite speed of light into account, small objects would appear rotated, rather than flattened, as they flash by. He wrote before black holes were understood, which is a pity in view of the wonderful use that he could undoubtedly have made of them in his stories. The inevitability and the nature of the cosmic singularities that occur in black holes and the big bang lead us now to doubt the oscillating nature of the closed universe. But Gamow's prophetic judgements concerning cosmology and the origin of the universe have profoundly stood the test of time. It remains undecided, as Gamow was undecided, whether the universe is spatially open or closed; but the steady-state has not survived, as Gamow had predicted, and his big-bang picture of the universe's origin has convincingly won the day.

ROGER PENROSE
October 1992

220

MATHEMATICS, PHYSICS, COMPUTERS AND CONSCIOUS UNDERSTANDING

The dancer and the dance. Trinity Jameson quatercentenary symposium (1992) Eds E. Sagarra and M. Sagarra. Trinity College, Dublin, 1993

Professor Roger Penrose

Mathematics, Physics, Computers and Conscious Understanding

It is certainly a great pleasure and a great honour for me to be able to address the 400th anniversary of the College, particularly because there are several people who have been associated with Trinity College who have had a very profound influence on my own work. Specifically I refer to William Rowan Hamilton, to Fitzgerald and to Synge — and when I refer to Fitzgerald and Synge I mean, of course, the mathematical physicists and not the writers.

George Francis Fitzgerald was influential for me because of his first steps towards relativity in the Fitzgerald contraction, and John Layton Synge because of his geometrical approach to relativity which I certainly found very congenial and which has had a very great influence on the way in which my own ideas have developed. I am indeed pleased that he is still around at the age of ninety-five. All these people have contributed to the understanding of the physical world in terms of mathematics, and it is a profound thing to see this very close unity between mathematics and physical understanding. Moreover it is often not just mathematics, but sophisticated mathematics. The relations that have been discovered between mathematics and the physical world have really had a two-way influence. Mathematics has been essential to the development of theoretical physics and hence to the overall development of physics, while, in its turn, physics itself has had a tremendous impact on mathematics. We have seen, particularly in recent years, yet one more wave of this kind of activity which, of course, goes back to Newton and even earlier.

The issue of understanding the world in terms of mathematics has, in modern times, obtained a new perspective. This has arisen in relation to the development of high-speed computers and the attendant computational techniques. Certainly, this has been one of the major revolutions of this century, particularly the second half of the century. It has also brought home — or made more pertinent and imminent — the issue of the relationship between the mind and computation. A not unreasonable case can be made that perhaps all that we do with our minds is some form of computing. If that is the case, then at some point in the development of computers — since computers are in most ways far more effective at computing (or calculating) then we are ourselves — they will eventually be able to do better than we can whatever operations we perform with our minds.

There is one particular point that I want to stress, namely that there is a profound difference between mathematics and computing. Or perhaps I should say computing is just part of mathematics. It is very misleading to think of mathematics merely in terms of computing. No doubt the development of very powerful computers and familiarity with them has tended to make people feel that perhaps this is all there is to mathematics. To illustrate the difference between mathematics and computing, let me mention a mathematical joke which — dare I say it — has something of an Irish flavour to it. The joke (which must be told by a mathematician) goes as follows: 'There are three kinds of mathematician, those who can count and those who cannot'. I think that it should be clear from that joke that I am counting myself amongst the mathematicians who cannot count, or, to put it a little less embarrassingly, those who cannot do arithmetic very well. It does perhaps divide mathematicians into two categories — there are certainly some very eminent mathematicians, perhaps the most eminent of all, who were very good at calculating and I mention Gauss in particular. But there are some eminent mathematicians who are not good at this.

Mathematics is something very different from calculating. Mathematics is about concepts, and it is about understanding. Calculation can be extremely valuable and often serves to substitute for understanding or to enhance understanding. One can often use one's understanding in profound ways, to formulate specific calculations — or what one calls algorithms — for solving problems. This is often an aim in mathematics, to analyse some area of mathematics with an aim to finding a procedure that reduces it to calculation. But this act of reducing it to calculation is more the kind of thing that mathematics is about, rather than the calculation itself. Often one takes the view that a problem is solved once one has reduced it to an algorithm.

What do I mean by an algorithm? I should say that algorithms are basically what computers do. There is a slight technicality involved in the fact that, strictly speaking, an algorithm may require a computer to have an unlimited store. But that is, I think, regarded generally as a technicality. Essentially, algorithms are what computers do. The idea was really first made precise by Alan Turing who also has Dublin connections, as I understand it, through his mother who was a Stoney — a relative of George Johnstone Stoney, the Dublin physicist who named the electron and first measured its charge, and who also, if I am not wrong, was connected with Fitzgerald. But here I want to refer more particularly to the famous theorem of Kurt Gödel, published in 1931, which I believe underlines the distinction between understanding and calculation. It was really Turing who made clear that an essential part of what Gödel was doing was basically concerned with calculation. Gödel was thinking more in terms of what are called formal systems in mathematics, but let us address the issues in terms of what can be done in principle by a computer. What Gödel showed, in essence, was that mathematical understanding and belief cannot be reduced to calculation, or to some set of rules. Gödel showed that if we believe such a set of rules — if we *know* and believe those rules — then that very knowledge provides us with a new rule not contained in the original rules themselves. Thus the rules could not have encapsulated all of our mathematical understanding or belief. Therefore there must be something in our understanding which goes beyond mere calculation or the following of rules. This being so in a mathematical context (and of course that is what Gödel was concerned with), it really opens the floodgates, because if there is something that we do with our minds which cannot be reduced to calculation in one area — mathematics — then we must expect that this will also be true in other areas of thought. Although the area of mathematical thought is where the theorems are, the implication is that there may well be, in *all* our (at least conscious) thinking, something that cannot be reduced to computation.

This is the point of view that I want to put forward and emphasize. I should say that it is a controversial view and there are some people who strongly hold to contrary views. I am trying to persuade you that there is something in our understanding which is fundamentally different from computation. I think that this issue is also important in other contexts, such as in teaching in schools, because I think there is a trend, certainly in England, to say that you should teach more facts, teach more rules, at the expense perhaps of the understanding which underlies these rules. Understanding is different from the following of rules. Understanding is something which cannot be reduced to rules. Understanding is a much more valuable thing than the mere following of rules.

Let me give you a very simple illustration of the difference between understanding and rule following. If you multiply two numbers together, say five times seven, it is the same thing as multiplying them in the opposite direction, seven times five. You can teach people these things. You could teach them case-by-case: eight times nine is the same as nine times eight, and so on. Or you tell them that there is a general rule that **a x b** is **b x a**. But that is different from giving the understanding of *why* **a x b** is the same as **b x a**. I have been referring to natural numbers. You have to be careful with the word 'number' because one of the things that Hamilton was famous for was introducing numbers in which **a x b** was *not* **b x a**. But those were not ordinary numbers, and I am talking about ordinary natural numbers here. That is important; the ordinary natural numbers are 0, 1, 2, 3, 4, 5, 6, and so on — those things — and the fact that you *know* what I mean when I say 'those things' is important. This actually is a critical point in the whole argument. You cannot merely, by any set of fixed rules, explain completely what one means by the natural numbers. Somehow that concept is accessible to us, and by means of rough descriptions we are able to 'latch on' to this number concept. But you cannot just give a set of rules to define the natural numbers. How do you see that **a x b** is **b x a** for natural numbers? Well, you imagine a rectangular array and you read it off by rows or columns; the first gives you **a x b** and the other **b x a**. Since it is the same rectangular array, you can see that it must be the same thing whichever way you multiply the numbers. And this gives you *belief* in the rules. It is not just knowing the rules, and knowing how to follow them. You can put rules on a computer, you can make it say **a x b** is the same as **b x a**, but it doesn't have the understanding that they *must* be the same (if you are talking about ordinary numbers 0, 1, 2, 3, 4, etc). It is this belief by understanding — and when I say belief, I don't mean blind faith, I mean genuine understanding — this is the thing which you cannot encapsulate in rules.

What Gödel really showed us was that if you have any set of rules which are meant to describe arithmetic, and if you believe that these rules give you only true facts (in particular, if you believe that these rules do not prove that $2=3$) then that very belief and understanding — of the *consistency* of those rules — can be coded into another rule which you must therefore also believe in. But this new rule is something which cannot be obtained from the rules that you started with. This is what Gödel showed: that there is always something more no matter what reliable rules you have. If you believe in those rules, then you can see how to get beyond them and obtain another rule which lies outside the rules you started with.

Now there are many areas where computational activity is very important in current research, say in artificial intelligence, where one is trying to imitate the behaviour of the mind by computation. The aforementioned considerations show that at least some kinds of thinking cannot be reduced to computional rules, i.e. they cannot be properly simulated on a computer.

When I use the term 'computer' here, I should make it clear that this includes parallel computers as well as the more familiar serial ones. Often people put faith into what are called massively parallel systems. But I should stress that this is not anything essentially different. Parallel systems perform operations which you could also perform on a serial computer if you wanted to. It is just that parallel ones may be more efficient for certain types of calculation; they may be much more efficient, but they do not in principle do things that you cannot do with an ordinary serial computer. Another thing which is often referred to is 'top-down' as opposed to 'bottom-up' architecture. Top-down means that you put in your rules, which are specific rules to solve particular problems. Or you may have a bottom-up architecture where you put in systems which improve themselves; they may have a rough kind of organization — artificial neural networks being one example (sometimes referred to as connectionist philosophy). Your computational system might have some kind of 'heuristics' — value rules which are not necessarily meant to solve the problem without error, but which act as a guide towards solving the problem. And you might have learning systems which improve themselves, you might have random elements incorporated, you might have chaotic systems, where these systems behave in an apparently random way when they are not perhaps, strictly-speaking, random. You might involve elements from the outside world, but all these things I am including in what I mean by an algorithm, by which I mean things you can do with a modern general purpose computer. Artificial neural networks, for example, are often simulated on an ordinary general purpose computer. And if you can put something on an ordinary general purpose computer, then it is included in what I mean by calculational or algorithmic. And this would also apply to the environment, provided that the environment is something which you can simulate.

The argument I am trying to make here is that there must be things about the way the physical world behaves which go beyond computation. The reason I believe this goes back to the Gödel theorem, as I have described, which tells us that there are things in the way our minds act which go beyond computation. I do believe that we must take a scientific attitude to what is going on inside our heads — that it is some kind of sophisticated physics that is going on — but it is a sophisticated physics which goes beyond computation. That there is, in physics, something which goes beyond computation is precisely the point I am trying to make. If the environment could not be simulated computationally, that would be just another instance of this.

I think that one way of looking at what the Gödel theorem is, is that it tells us that conscious understanding cannot be reduced to a set of rules. But it might be that *un*conscious understanding is governed by some set of rules that we do not know and that we do not understand. It might of course be that such things guide mathematicians to their inspiration; it might be that some unconscious action is what underlies their intuitions; but it does not seem to me that you can say that this is what underlies their *beliefs* because their beliefs are things that are shared by other mathematicians. Mathematical beliefs can be explained, they are very rigorous things, and it seems to me that this is something for which your consciousness is needed. In other words, what you decide you believe in, in mathematics, must be a conscious act. Although there may be some horrendous, perhaps inaccurate, calculational process going on about that we are not aware of, that does not explain our conscious beliefs.

Another thing Gödel's theorems tells us is that there are things in mathematics which go beyond computation; there are sets of problems which you could not ask a computer to solve and be sure to get the answers. Of course some might say: well how do we know that *people* could give us the answer? Gödel's theorem does not tell us that, but what it does tell us is that those insights which are accessible to our understanding — to our conscious understanding — are things which go beyond computation, so although we may not be able to solve all the problems, the insights which are available to us for solving such problems are things which go beyond computation. Now if it is not computation which is going on in our heads, what can it be? Here one gets more speculative. My own belief is that we have to look somewhere in physics where action of a kind that we don't yet understand is going on. I mean this at a very fundamental level. This is because classical physics is something that you could in principle simulate computationally — and quantum physics, if we are talking just of things going on at the level of atoms and molecules where the Schrödinger equation governs behaviour, is also computational. (Schrödinger lived a good part of his life in Dublin, so here Dublin comes back again!) Classical physics governs what happens at the large scale and quantum physics at the small scale. But there is, it seems to me, a big mystery in how the two tie together. It is in the tying of the two levels together that it seems likely that we must find a new type of physics where these non-computational ingredients will come in. Of course this is speculative, but what I am speaking about here is how we might perceive future developments in physics. This is a place where I believe some advances are likely to come.

I would like to finish by pointing out what I believe to be some dangers in the development of computers. The danger is not, as I see it, that computers will someday replace us. (It might be that some device *not* based purely on computational systems could eventually replace us. I have nothing to say on that one.) I think the danger is different. One danger is that we lose our pride by thinking that, since computers are objects which certainly do some things very much better than we do, they are our superiors. There are many other dangers, different kind of dangers, involved in computer technology. There is the possibility, for instance, of computer fraud, something which is quite well publicised, where people take advantage of computers — their understanding and knowledge of computers — to defraud some company of some person. There also is the intrusion into our privacy, which computers now make much easier than it used to be. There is the possibility of espionage, using computational systems and, I think, of very subtle kinds of sabotage. I want to mention one particular possibility which I have never actually seen suggested. Many of us are well aware of computer viruses. They are hidden parts of computer programmes that are introduced by some disreputable people — maybe disgruntled because they have lost their jobs or something. They insert some nasty sequence into a computer programme which gets around from computer to computer and, at a certain date, will destroy everybody's computer system which happens to have had contact with the virus. It seems to me that these things could be used in much more subtle ways. It is something we have to be on the look out for if commercial concerns, perhaps even governments, were to take advantage of the possibility that one might be able to influence the action of a computer, or of all computers that might be concerned with a particular problem. They may be able to produce results which look as though they are correct but in fact are not correct. Now I do not know whether anything like this has actually been used, but it seems to me to be something which might be done. It is something that in my view one has to guard against by making sure that we do not rely too much on computers. In other words we have to have systems where we can check up on what they are actually doing by methods which are *not* done by general purpose computers.

221

AN EXAMPLE OF INDETERMINACY IN THE TIME-DEVELOPMENT OF "ALREADY UNIFIED FIELD THEORY": A COLLISION BETWEEN ELECTROMAGNETIC PLANE WAVES

Directions in general relativity, Vol. 1. Eds B. L. Hu, M. P. Ryan Jr., and C. V. Vishveshwara, 288–96. Cambridge University Press, 1993

This paper was written at a time when John Wheeler, then at Princeton, had been promoting the striking idea of what he called 'geometrodynamics', involving an 'already unified' geometrical space-time theory, according to which the space-time metric was to determine all other fields. In particular this applied to electromagnetism, where Charlie Misner and Wheeler had found a way (actually reproducing work by G. Y. Rainich, done in 1925) whereby the information of an electromagnetic field could, in general circumstances, be effectively reconstructed from its 'imprint' on the metric field. In the present article, I had shown that these equations could not always be deterministic, but the manuscript was mislaid for a while, and had rather lost its interest by the time it was found. But much later I revived it to honour Misner's 60th birthday. For myself, this paper stimulated my particular later fascination with colliding plane waves, which showed itself in Chapters 21 and 44.

340 An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves. *Directions in general relativity*, Vol. 1. Eds B. L. Hu, M. P. Ryan Jr., and C. V. Vishveshwara, 288–96. 1993

An Example of Indeterminacy in the Time-development of "Already Unified Field Theory": A Collision between Electromagnetic Plane Waves

R. Penrose[*†]
Palmer Physical Laboratory
Princeton, New Jersey.

Abstract

An example is presented which points to a certain basic difficulty in the "already unified" approach to unified field theory. It is shown that one can construct a pair of solutions of the combined Einstein-Maxwell equations for which the two space-times are identical in the neighbourhood of an initial space-like hypersurface (and in fact they may also be identical at all earlier times), but the time-development of the equations leads to space-times which are essentially different in their futures. The construction of such examples requires the electromagnetic field to be null (or zero) in some regions. The example given here represents a collision between two gravitational-electromagnetic waves.

1. Introductory preamble

This paper was written in late 1959 or early 1960, while I was at Princeton University in the early part of my research career in general relativity. It was at a time when I knew Charlie Misner best, since he was also in Princeton then, and I learnt a great deal from him about issues of general relativity, such as the initial value problem etc. As far as I can recall, it was discussions with him, and also with John Wheeler, that led to the ideas described in this paper.

I had completed the paper, and gave it to John Wheeler for his comments. Unfortunately, unforseen circumstances intervened, and it was not until several months later

---

[*] NATO Science Visiting Fellow

[†] present address (1993): Mathematical Institute, Oxford, U.K.

that the paper resurfaced, at which time my own interests had moved elsewhere. The celebration of Charlie's 60th birthday seemed an ideal occasion on which to resurrect the paper, and I searched through old files in order to locate it. Fortunately I located almost all of it, but unfortunately page 3 of the typescript was nowhere to be found. I have done my best to reconstruct a substitute (positioned between the two dark horizontal lines in the article), although I am not really sure what equations (4) and (5) were. I do not think that this is of too great consequence, however. Apart from this insertion, and the completion of some references, I have left the original article completely unchanged.

As a final comment, in these introductory remarks, I should make reference to some work that appeared in two later papers[13,14]. Evidently, in 1959/1960, I was not aware that plane waves do not possess global Cauchy hypersurfaces[13], though I had some partial conception of the singularities that were likely to arise[14]. I hope that this paper, old though it is, is a worthy tribute to Charlie Misner's many original contributions to general relativity.

2. Introduction

According to the reformulation of the combined Einstein-Maxwell theory put forward by Rainich[1] and by Misner and Wheeler[2] it is possible, under normal circumstances, to infer from the metric alone, all that is necessary for the physics of gravitation plus source-free electromagnetism. However, it will be shown here that in certain special cases the future space-time (*within* the domain of dependence) cannot be predicted from a knowledge of the metric in the neighbourhood of an initial space-like hypersurface alone, nor sometimes even from the whole past space-time metric. It follows, therefore, that in the "already unified" theory, no amount of allowable data on an initial hypersurface can be sufficient *in all cases* to determine the future uniquely. This situation is in sharp contrast with the analogous (Cauchy) problem for conventional general relativity, as the work of Foures-Bruhat[3] shows.

In order to present an example which has this indeterminacy property, it will be convenient first to discuss some rigorous solutions of th Einstein-Maxwell equations which are the curved-space analogues of the plane-wave solutions of Maxwell's equations.

3. Plane Waves

A metric representing the general plane-polarized gravitational-electromagnetic plane wave has been given by Rosen[4]:

$$ds^2 = e^{\alpha}(dt^2 - dx^2) - e^{\gamma+\beta}dy^2 - e^{\gamma-\beta}dz^2 \tag{1}$$

where α, β, γ are functions of $t - x$. The condition

$$2\gamma'' + \gamma'^2 + \beta'^2 - 2\alpha'\beta' \leq 0 \tag{2}$$

must also be imposed to ensure that the time-time component of the Ricci tensor is non-positive. With this condition satisfied, a Maxwell field can be constructed which satisfies, with the metric, the combined Einstein-Maxwell equations. The Maxwell field so constructed turns out to be necessarily null; that is, both its invariants vanish.

The necessary and sufficient condition for the Maxwell field to *vanish* for the metric (1) is that the equality sign in (2) should hold. The metric (1)

(reconstruction of essentials of missing page)

represents an entirely electromagnetic wave (i.e. vanishing Weyl tensor, cf. (7)) if

$$\beta'' + \beta'\gamma' - \beta'\alpha' = 0. \tag{3}$$

We could, if desired, arrange that $\alpha = 0$ in (1) by rescaling the $t = x$ coordinate. However, α is retained here for later convenience.

There is also an alternative form[7,8] for this metric, namely

$$ds^2 = dudv + h_{ij}(u)x_i x_j du^2 - dx_i dx_i, \tag{4}$$

where we require

$$h_{ii} \geq 0 \tag{5}$$

to ensure that the time-time Ricci component is non-positive.

Bondi, Pirani and Robinson[9], following Robinson[7], showed (in the purely gravitational case) how the metric (1) could be used to describe a

---

non-singular "sandwich wave" which separates two regions of flat (Minkowskian) space. This situation is shown in Fig. 1. The space is exactly flat before the wave arrives and is again exactly flat after it has departed. The curved region is bounded by two (null) hyperplanes. Bondi, Pirani and Robinson considered only the purely gravitational case in their paper, but Rosen's metric (and also the metric (4)) allows the construction of similar sandwich waves where electromagnetic field is present in the curved region. A simple way of constructing such an electromagnetic sandwich wave is to choose $\beta \equiv 0$, so that (3) is automatically satisfied, choose $\gamma \equiv 2(t-x)$ and $\alpha' \equiv \Omega + 1$ where $\Omega \geq 0$ is a function of $t-x$ which is sufficiently smooth and which vanishes outside a certain region. Thus (2) is satisfied and satisfied with equality outside this region. Also (3) is satisfied everywhere so that outside this region the space-time is flat. The metric so chosen does not become singular, but the space-time so defined is not complete. To complete it, flat pieces with Minkowskian coordinates may be adjoined so as to overlap the flat regions of the metric (1) in the manner of Bondi, Pirani and Robinson. The transformations to Minkowskian coordinates take the form

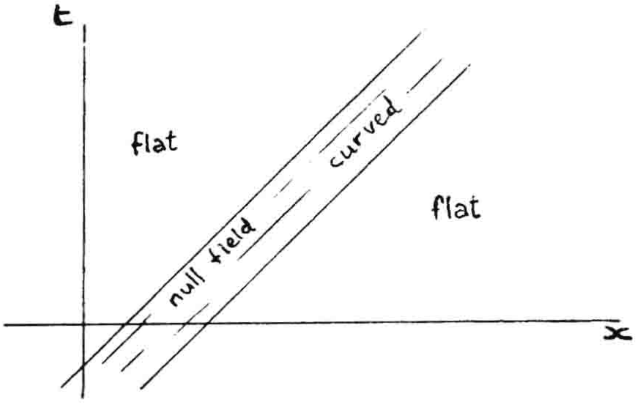


Fig. 1. Space-time picture of the sandwich wave.

$$t - x = \log(\tau - \xi),\ t + x = \frac{\tau^2 - \xi^2 - \eta^2 - \zeta^2}{a(\tau - \xi)},\ y = \frac{\eta}{\tau - \xi},\ z = \frac{\zeta}{\tau - \xi}, \tag{6}$$

where a is a suitable constant. The curved regions are bounded by null hyperplanes ($\tau - \xi =$ constant) in the Minkowski spaces.

The tensor structure of such waves will be considered in Sec. IV.

4. Collision between Sandwich Waves

Consider the following situation. Suppose two such electromagnetic sandwich waves $\mathcal{A}$ and $\mathcal{B}$ approach each other from infinity. Let S be a space-like (initial) hypersurface which intersects these waves before they collide and let A and B be the respective intersections of the waves $\mathcal{A}$ and $\mathcal{B}$ with S. (Fig. 2) It is clear that *whatever* initial value conditions are imposed on the *metric* in the neighbourhood of S, they will still be satisfied if two *independent* constant duality rotations[10] are applied to the electromagnetic fields in the neighbourhoods of A and B respectively. The space-time in the neighbourhood of S is unaffected by such duality rotations. However, at a later time, after the two waves have collided and are interfering, the space-time *will* be affected if *different* duality rotations of the electromagnetic field are applied to each of the two waves.

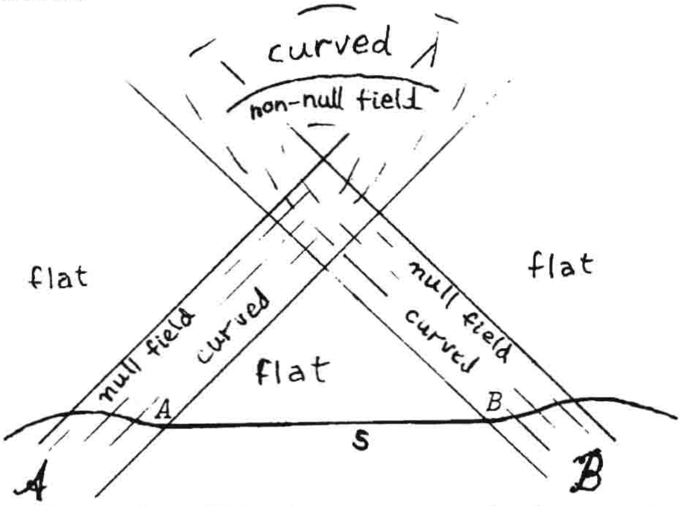


Fig. 2. Space-time picture of a collision between two sandwich waves $\mathcal{A}$ and $\mathcal{B}$ which approach each other from infinity. Before colliding the curved regions are bounded by flat characteristic hypersurfaces and the space-time between the waves is flat. When the waves recede again from each other, their non-linear interference causes the space-time between them to be curved and the chracteristic hypersurfaces which bound then on the outside also to be curved. The initial space-like hypersurface S may be chosen to be flat between the waves, but if so, it is necessarily non-flat outside the waves owing to the warping effect of the curvature in $\mathcal{A}$ and $\mathcal{B}$.

One way of seeing this is to examine the Maxwell tensor in the region where the waves first overlap. Assuming the fields to be initially weak, the linear addition of the two fields will at first dominate any non-linear effects. But it can be easily seen that the sum of two differently duality rotated Maxwell tensors is not a duality rotation of the sum of the two original Maxwell tensors. Hence, if different duality rotations are applied to the two waves, the result must show up in the metric after they meet. (Only duality rotations do not affect the metric.) Also, in order to see that the space-time can be continued into the future, at least for a finite[11] time, it is sufficient to consider a space-like hypersurface passing through the (space-like 2-plane) region of first meeting of the waves. The Maxwell field and metric satisfies the required initial value conditions for this hypersurface and can therefore be continued for a finite time into the future (see Foures-Bruhat[3]).

Thus we have an example of two space-times which are identical in the neighbourhood of a certain space-like surface — and in fact they are also identical at all earlier times — but which are essentially different in their later time-developments. Both space-times admit Maxwell-type fields which satisfy, with the metric, the combined Einstein-Maxwell equations.

One might ask the question; how is this example compatible with the analysis given by Rainich and by Misner and Wheeler? All the necessary information about the Maxwell field ought to be contained in the metric, with no room for later indeterminacies. One answer to this question is that the Maxwell field in the whole initial region is *null* (i.e. both its invariants vanish). This is the case not treated by these authors. However, it seems to be possible to modify this example so that the *non-zero* Maxwell field is nowhere exactly null. This would seem to be the case if, instead of plane waves, radiation from two finite sources at a large distance apart had been used. The essential feature of such an example is that two regions of non-zero field are separated by a region where the Maxwell field vanishes. This region of zero field can be thought of as a special case of a null field, so it falls outside the scope of Rainich and Misner-Wheeler treatments.

It is of interest, however, to give an example which does *not* depend on two regions of non-zero Maxwell field being separated by a region of *zero* Maxwell field, but nevertheless indeterminacies arise in the future development of the space-time. Such an example (necessarily null in some parts of the space-time) can be given in terms of the plane wave collision described above, but it will be necessary first to examine the structure of the plane waves more closely. The plane waves exhibit a curious freedom with regard to duality rotations of the Maxwell field, which can only occur in the null case.

5. Tensor Structure of the Plane Waves

The Riemann tensor and Maxwell field tensor structure of the plane-waves can be inferred from some results given in Penrose[6]. The results are given there in a spinor form but they can easily be translated into a tensor form. At any point of a space-time in which the Einstein-Maxwell equations are satisfied, the Riemann tensor can be decomposed as follows[2]:

$$R_{\mu\nu\rho\sigma} = C_{\mu\nu\rho\sigma} + F_{\mu\nu}F_{\rho\sigma} + F^*_{\mu\nu}F^*_{\rho\sigma} \tag{7}$$

where $F_{\mu\nu}$ is the Maxwell field tensor defined in suitable "geometric" units (velocity of light = 1, gravitational constant = $1/(4\pi)$) and $F^*_{\mu\nu}$ its dual, defined by

$$F^*_{\mu\nu} = \frac{1}{2}(-g)^{\frac{1}{2}} F^{\alpha\beta}\epsilon_{\alpha\beta\mu\nu} \tag{8}$$

The tensor $C_{\mu\nu\rho\sigma}$ is Weyl's conformal tensor. It satisfies $C^{\mu}_{\nu\mu\sigma} = 0$ and can be thought of as describing the *purely gravitational* part of the curvature, the electromagnetic part having been subtracted out. Its dual $C^*_{\mu\nu\rho\sigma}$ is defined by

$$C^*_{\mu\nu\rho\sigma} = \frac{1}{2}(-g)^{\frac{1}{2}} C^{\alpha\beta}_{\mu\nu}\epsilon_{\alpha\beta\mu\nu} = \frac{1}{2}(-g)^{\frac{1}{2}} C^{\alpha\beta}_{\rho\sigma}\epsilon_{\alpha\beta\mu\nu} \tag{9}$$

the right and left duals being equal.

An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves. *Directions in general relativity*, Vol. 1. Eds B. L. Hu, M. P. Ryan Jr., and C. V. Vishveshwara, 288–96. 1993

Consider the case of a plane-wave and choose, for convenience, a point O at which neithre the Maxwell field tensor $F_{\mu\nu}(O)$ nor the Weyl tensor $C_{\mu\nu\rho\sigma}(O)$ vanishes. Let X be any point of the manifold and let OX be a geodesic joining O to X. Let x^{μ} be a vector defined at O which is tangent to OX at O and whose length is the length fo OX (interpreted suitably if OX is null). Then the Maxwell field tensor and Weyl tensor at the point X are given (comparing the tensors at the two points by parallel transfer along OX) by

$$F_{\mu\nu}(X) = a(F_{\mu\nu}(O)\cos\theta - F^{*}_{\mu\nu}(O)\sin\theta) \tag{10}$$

$$C_{\mu\nu\rho\sigma}(X) = b(C_{\mu\nu\rho\sigma}(O)\cos\phi - C^{*}_{\mu\nu\rho\sigma}(O)\sin\phi), \tag{11}$$

a, b, θ, ϕ being real functions of $x^{\lambda}p_{\lambda}$ where p_{λ} is a null vector giving the direction of motion of the wave. The functions b determine the amplitudes and the functions θ, ϕ determine the planes of polarization of, respectively, the purely electromagnetic and the purely gravitational parts of the wave. Each of these four real functions may be chosen arbitrarily (apart from considerations, possibly, of continuity o differentiability) and these four functions determine the space-time[12]. Thus, to construct a sandwich wave, we merely choose for a and b, two functions which vanish outside a certain finite interval. If, as was the case with the particular sandwich wave constructed in II, the Weyl tensor vanishes everywhere we have $b \equiv 0$, so the point O need only be chosen where the Maxwell field does not vanish.

Only a, b and ϕ affect the geometry of the space-time. A change in the function $\theta(x^{\lambda}/p_{\lambda})$ leads merely to a duality rotation of the Maxwell field at each point. This duality rotation is constant along each hypersurface $x^{\lambda}p_{\lambda}$ = constant, (these are null hyperplanes) but it is otherwise arbitrary. Thus a (continuous) duality rotation Θ may be performed which affects the trailing half of a sandwich wave only, the Maxwell field in the leading half being unaffected.

It may be mentioned here, in possing, that these duality rotations correspond to *actual* space rotations of the field tensor (rotation of plane of polarization). This is a characteristic feature of the null field and it shows why no scalar "complexion" can be defined for the null case. It may also be mentioned here that whereas the Maxwell field is null for these plane waves, so also, in a sense, is the Weyl tensor, that is (see Bondi *et al*[9], Penrose[6]) it is of Petrov type II with vanishing invariants.

6. Indeterminacy with Non-zero Field

We are now in a position to consider again the space-time of Fig. 2 which represents a collision between plane waves. However, let us this time choose as the initial hypersurface a space-like hypersurface S' which intersects the two wave regions after they have collided but before the waves have completely overlapped. The situation is as shown in Fig. 3. Let A' and B' be the respective intersections of the waves $\mathcal{A}$ and

$\mathcal{B}$ with S' (A' and B' now overlap) and let C' be the intersection of a trailing portion $\mathcal{C}$ of the wave $\mathcal{A}$ with S'. Choose $\mathcal{C}$ so that, as in Section IV, a duality rotation Θ of the Maxwell field in $\mathcal{A}$ (prior to collision) may be performed which affects only the field in $\mathcal{C}$ and so does not affect the field in the rest of $\mathcal{A}$. Make sure S' is chosen so that B' and C' are separated by a region of A'. Now the duality rotation Θ does not affect the space-time in the neighbourhood of S' but, as in the previous example, it *will* affect the later space-time when $\mathcal{C}$ and $\mathcal{B}$ have overlapped.

Thus, an example is provided with the proprety that the future space-time is not determined by the space-time in the neighbourhood of the initial hypersurface, but nevertheless no two regions of non-zero Maxwell field on this hypersurface are entirely separated by a region of zero Maxwell field. The essential feature of this example is that the regions B' and C' are separated by a region of *null* Maxwell field. It is also possible to arrange that the Maxwell field be *nowhere* zero on S', if desired, by letting the amplitudes of the waves tail off to infinity behind them.

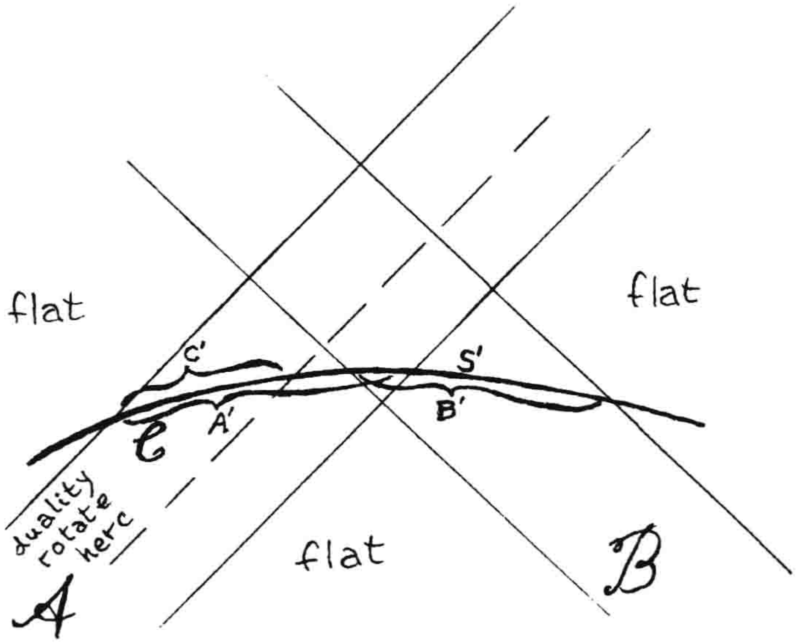


Fig. 3 Sandwich wave collision in more detail and with later initial hypersurface

7. Acknowledgements

I am most grateful to Dr I. Robinson, Professor J.A. Wheeler and Dr W. Kundt for valuable discussions concerning various aspects of this paper. I am grateful also to K. P. Tod for some assistance in connection with the missing p.3.

[1] G.Y. Rainich, Trans. Am. Math. Soc. **27**, 106 (1925).

[2] C.W. Misner and J.A. Wheeler, Ann. Phys. **2**, 525 (1957).

[3] Y. Foures-Bruhat, J. RAtional Mech. Anal. **5**, 951 (1956).

[4] N. Rosen, Phys. Z. Sowjet. **12**, 366 (1937).

[5] H. Takeno, Tensor **8**, 59 (1958).

[6] R. Penrose, Ann. Phys. **10**, 171-201 (1960).

[7] I. Robinson, Report to the Royamont Conference (1959).

[8] I. Robinson, Report to the Eddington Group, Cambridge, 1956, H W Brinkman, Proc. Natl. Acad. Sci. (U.S.) **9** (1923), 1.

[9] H. Bondi, F.A.E. Pirani and I. Robinson, Proc. Roy. Soc. London **A251**, 519 (1959).

[10] A duality rotation (see Misner and Wheeler[2]) of the Maxwell field tensor $F_{\mu\nu}$ is given by $F_{\mu\nu} \to F_{\mu\nu}\cos\theta - F^{*}_{\mu\nu}\sin\theta$ where $F^{*}_{\mu\nu}$ is the dual of $F_{\mu\nu}$ (see eq. (8)). Thus, in particular, a duality rotation through 90° interchanges (apart from a sign) the electric and magnetic fields. The Maxwell stress energy tensor is invariant under any duality rotation (and under no other transformation) of $F_{\mu\nu}$.

[11] It will be shown in another paper, however, that after the collision the wave fronts necessarily become eventually singular. It seems likely that the space-time may eventually become singular also.

[12] In Penrose[6] these functions were taken to be analytic. However, this is inessential since any continuous function of a real variable can be smoothly approximated arbitrarily closely by an analytic function. Alternatively, C^{∞} piecewise analytic functions may be used. Such functions have all the generality that is required here.

[13] R. Penrose, Revs. Mod. Phys. **37** 215-220 (1965)

[14] K.A. Khan and R. Penrose, Nature **229** 185-186 (1971)

222

ORTHOGONALITY OF GENERAL SPIN STATES

Twistor Newsletter **36** (1993)

The problem considered here was essentially one that had concerned me since my early Cambridge days in the 1950s, but it has relevance to the quantum geometry of spin, such as that described in Chapters 227 and 291. Some of the geometry that comes about in the present work led to the results of Chapters 224 and 286. [Added in proof, June 2010] Maciej Dunajski has just informed me of a much earlier work on this topic, by C. F. Russell (*Proc. Lond. Math. Soc.*, 1907), On the Geometrical Interpretation of Apolar Binary Forms.

Orthogonality of General Spin States

E. Majorana (*Nuovo Cimento* $\underset{\sim}{9}$ (1932) 43–50) described the general spin state, for spin $\frac{n}{2}$, (up to proportionality) in terms of an unordered set of n points on the sphere – perhaps with coincidences between them, but then multiplicities are counted. Although he did not phrase his argument in these terms, his result essentially expresses the canonical decomposition of a general symmetric spinor of valence n:

$$\psi_{\underbrace{AB\cdots N}_{n}} = \alpha_{(A}\beta_{B}\cdots\nu_{N)} \, .$$

Here $\psi_{AB\cdots N}$ describes the spin state, and its n principal null directions, represented as n points on the Riemann sphere S, provide Majorana's set (cf. Penrose & Rindler, *Spinors & Space-Time*, Vol. 1 (1984) p. 162; R.P. in *300 years of Gravity* (eds. Hawking & Israel) C.U.P. (1987); R.P. in *The Emperor's New Mind*, O.U.P (1989), Fig. 6.29).

I shall be concerned, here, with the question of the geometrical interpretation of *orthogonality* between two states of spin $n/2$. In spinor terms this can be written

$$\psi_{AB\ldots N}\, t^{AA'}t^{BB'}\cdots t^{NN'}\, \overline{\phi}_{A'B'\ldots N'} = 0$$

where t^a is the timelike (unit) vector (in 4-space), with respect to which the states are all taken to be stationary. We can write this last relation as

$$\psi_{AB\cdots N}\, \varphi^{AB\cdots N} = 0$$

where

$$\varphi^{AB\cdots N} = t^{AA'}t^{BB'}\cdots t^{NN'}\, \overline{\phi}_{A'B'\cdots N'} \, .$$

Geometrically, the n Majorana points on S defined by $\varphi_{AB\ldots N}$ are *antipodal* to the n points defined by

$\phi_{AB\ldots N}$ (see Spinors & Space-Time, Vol **1**). The relation between $\varphi_{AB\ldots N}$ and $\psi_{AB\ldots N}$ given by $\psi_{AB\ldots N}\,\varphi^{AB\ldots N}=0$ is called apolarity (cf. Grace & Young, The Algebra of Invariants C.U.P. (1903)). Apolarity represents a simpler geometrical condition for a pair of sets of n points on S than does orthogonality — not least because apolarity is conformally invariant on S (since there is no dependence on t^a) whereas orthogonality is only rotationally invariant. Then to pass from apolarity to orthogonality, we simply reflect one or other of the sets of points in the centre of S.

Nevertheless, apolarity is itself not an easy thing to express in purely geometrical terms, in the general case. In terms of components, the algebraic condition is

$$\psi_0\varphi_n - n\psi_1\varphi_{n-1} + \frac{n(n-1)}{2!}\psi_2\varphi_{n-2} - \ldots + (-1)^n\psi_n\varphi_0 = 0$$

where we can think of the n Majorana points of $\psi_{AB\ldots N}$ as defined on the Riemann sphere by the roots of

$$\psi_0 + \psi_1 z + \psi_2 z^2 + \ldots + \psi_n z^n$$

($\psi_0 = \psi_{00\ldots 0}$, $\psi_1 = \psi_{10\ldots 0}$, etc.) whereas orthogonality is given by

$$\psi_0\bar{\phi}_0 + n\psi_1\bar{\phi}_1 + \frac{n(n-1)}{2!}\psi_2\bar{\phi}_2 + \ldots + \psi_n\bar{\phi}_n = 0.$$

Let us denote by $P_1, \ldots, P_n$ the points on S defined by $\psi_{AB\ldots N}$, and by $Q_1, \ldots, Q_n$ those defined by $\varphi_{AB\cdots N}$. In a recent paper (R.P. "On Bell Nonlocality Without Probabilities, Some Curious Geometry" to be published in a CERN volume in honour of J.S. Bell; cf. also **J.** Zimba & R.P. "On Bell... ...More Curious Geometry," to be published: HPS Cambridge 1993), a number of special cases of apolarity are given:

① If all the P-points coincide then they are apolar to the Q-points iff at least one of the Q-points coincides with the n-fold P-point.

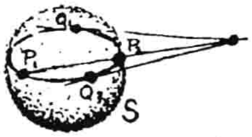
② If all but one of the P-points coincide, then the P- and Q-sets are apolar if stereographic projection from the multiple P-point sends the remaining P-point to the centroid of the stereographic projection of the Q-points.

③ If n is odd, then any set of P points on S is apolar to itself.

Let us now consider the general case.

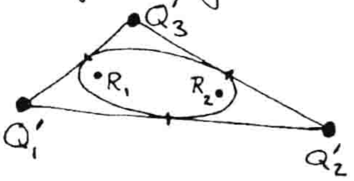
$\underline{n=1}$ Apolarity holds iff $P_1 = Q_1$

$\underline{\underline{n=2}}$ Apolarity holds iff P_1, P_2 separate Q_1, Q_2 harmonically — i.e. P_1, P_2, Q_1, Q_2 all lie on a circle on S and the line P_1P_2 meets the intersection of the tangents at Q_1 and at Q_2 to this circle (or limiting cases)



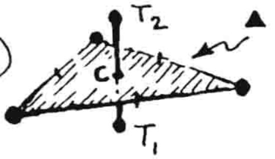
$\underline{\underline{n=3}}$ Stereographically project from P_3 (assumed simple — otherwise ② above); we get points P_1', P_2', Q_1', Q_2', Q_3' as the respective projections of P_1, P_2, Q_1, Q_2, Q_3.

Two ways of seeing the relation of P_1', P_2' to the triangle $Q_1'Q_2'Q_3'$ ($=\triangle$) that asserts apolarity are as follows.

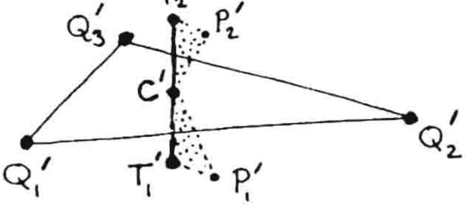


$\boxed{1}$ R_1, R_2 are the foci of the ellipse touching the sides of $\triangle$ at their mid-points. Then the P-points are apolar to the Q-points iff P_1', P_2' separate R_1, R_2 harmonically (in the above sense ($n=2$) — i.e. they separate harmonically on a circle through them).

$\boxed{2}$ Imagine a cardboard equilateral triangle ▲ with a matchstick through its centre C and perpendicular to its plane. The matchstick extends to points T_1, T_2 on either side of ▲ to a distance equal to its in-radius (i.e. $\frac{1}{2}$ its circumradius).



Now imagine that ▲ is held (in 3-space) so that its orthogonal projection to the plane coincides with △, scaling up or down as necessary (i.e. place ▲ so that, from a long way off, it "looks like" △). See where the points T_1, T_2, C project to; call them T_1', T_2', C'. The required apolarity condition between the P-points and Q-points is: $C'P_2'T_2'$ and $C'T_1'P_1'$ are similar.



Proofs: Exercise for the reader; hint: look at the accompanying article by MG E & RP (also, note that R_1, T_1', R_2, T_2' form a square with centre C').

General n ; inductive argument

Assume we already know a geometrical criterion for apolarity for two pairs of n-1 points. Can we find such a criterion for two pairs $P_1, \cdots, P_n$; $Q_1, \cdots, Q_n$ of sets of n points on S ?

(An) Answer (not very practical!) Fix one of the P-points, say P_n, and all the Q-points, and try to find those special sets of points $P_1, \cdots, P_{n-1}$ with $P_1, \cdots, P_{n-1}, P_n$ apolar to $Q_1, \cdots, Q_n$, for which $P_1 = P_2 = \cdots = P_{n-1}$. There are n-1 such places — call them $R_1, \cdots, R_{n-1}$ — these being the points on S, stereographic from which would yield points P_n' ; $Q_1', \cdots, Q_n'$ with P_n' the centroid of $Q_1', \cdots, Q_n'$ (by ② above). (Unfortunately I don't of a direct construction of $R_1, \cdots, R_{n-1}$.)

Apolarity between $P_1, \dots, P_n$ and $Q_1, \dots, Q_n$ is now simply the condition that $P_1, \dots, P_{n-1}$ and $R_1, \dots, R_{n-1}$ be apolar.

Proof Consider $\overset{n}{P}{}^{(A} X^B X^C \dots X^{N)} \overset{1}{Q}_{(A} \overset{2}{Q}_B \cdots \overset{n}{Q}_{N)} = 0$. suffices have now become numbers above

Solutions for X^A give $\overset{1}{R}{}^A, \dots, \overset{n-1}{R}{}^A$. Hence

$$\overset{n}{P}{}^A \overset{1}{Q}_{(A} \overset{2}{Q}_B \cdots \overset{n}{Q}_{N)} = \overset{1}{R}_{(B} \cdots \overset{n-1}{R}_{N)} \leftarrow \text{taking multiplying factor} = 1$$

Apolarity between $P_1, \dots, P_{n-1}$ and $R_1, \dots, R_{n-1}$ is

$$\overset{1}{P}{}^{(B} \dots \overset{n-1}{P}{}^{N)} \overset{1}{R}_{(B} \cdots \overset{n-1}{R}_{N)} = 0$$

which is the same as the required condition

$$\overset{n}{P}{}^{(A} \overset{1}{P}{}^B \dots \overset{n-1}{P}{}^{N)} \overset{1}{Q}_{(A} \overset{2}{Q}_B \cdots \overset{n}{Q}_{N)} = 0.$$

Q.E.D.

223

A SIMPLE OBSERVATION CONCERNING {2 2} VACUUMS

Twistor Newsletter **36** (1993)

See Chapter 240.

A Simple Observation Concerning {22} Vacuums

It is well known that every {22} vacuum ("type D"), with Weyl spinor

$$\Psi_{ABCD} = M r^{-3} \alpha_{(A}\alpha_{B}\beta_{C}\beta_{D)}$$

(where M is a constant and $\alpha_A \beta^A = 1$) possesses a Killing spinor — a valence-2 symmetric twistor! —

$$\chi_{AB} = r\,\alpha_{(A}\beta_{B)} \qquad \text{satisfying} \quad \nabla_{A'(A}\chi_{BC)} = 0.$$

(See Walker & Penrose, Comm. Math Phys **18** (1970) 265–74; also Spinors & Space-Time Vol. 2, p. 107.) In the Schwarzschild solution, r is the standard "radial coordinate", but in general r is a complex "radial" quantity.

I am not aware that anyone has pointed out the following simple but striking consequence:

Proposition Along every null geodesic, with parallelly propagated tangent spinor o^A, the null datum

$$\Psi_0 = \Psi_{ABCD}\, o^A o^B o^C o^D = C_{abcd}\, l^a m^b l^c m^d$$

has the precise form

$$\Psi_0 = \frac{K}{r^5}$$

where K is a complex constant depending on the choice of null geodesic (with its choice of scale for o^A).

Proof This is an immediate consequence of the constancy of $\chi_{AB}\, o^A o^B$ along the null geodesic ($o^A \bar{o}^{A'} \nabla_{AA'} (\chi_{BC}\, o^B o^C) = 0$) and the fact that $\Psi_0 = M r^{-3} (\alpha_A o^A)^2 (\beta_B o^B)^2$ and $\chi_{AB}\, o^A o^B = r (\alpha_A o^A)(\beta_B o^B)$, whence $\Psi_0 = M/(\chi_{AB}\, o^A o^B)^2 r^5$. Q.E.D.

Roger Penrose

224

THE ORTHOGRAPHIC IMAGE OF A REGULAR TETRAHEDRON

Twistor Newsletter **36** (1993). With M. G. Eastwood

358 The Orthographic image of a regular tetrahedron. *Twistor Newsletter* **36** (1993).
With M. G. Eastwood

The Orthographic Image of a Regular Tetrahedron

Michael Eastwood and Roger Penrose

Theorem Four points $\alpha, \beta, \gamma, \delta \in \mathbf{R}^2$ are the images of the vertices of a regular tetrahedron in $\mathbf{R}^3$ under orthogonal projection $\mathbf{R}^3 \to \mathbf{R}^2$ if and only if

$$(\alpha + \beta + \gamma + \delta)^2 = 4(\alpha^2 + \beta^2 + \gamma^2 + \delta^2) \tag{1}$$

where $\alpha, \beta, \gamma, \delta$ are regarded as complex numbers.

Proof. It is easy to check that (1) is translation invariant, so we may as well suppose that $\delta = 0$ and that the prospective regular tetrahedron has its corresponding vertex at the origin in $\mathbf{R}^3$. One possibility for the other three vertices is

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Taking the projection to be onto the last two coördinates we get, in this case,

$$\alpha = 1, \qquad \beta = i, \qquad \gamma = 1 + i$$

and it is easy to check that

$$(\alpha + \beta + \gamma)^2 = 4(\alpha^2 + \beta^2 + \gamma^2) \tag{2}$$

as required. The conformal orthogonal group may be double covered by the group of invertible 2×2 complex matrices of the form

$$\Lambda = \begin{pmatrix} a & -b \\ \bar{b} & \bar{a} \end{pmatrix}$$

acting by $X \mapsto \Lambda X \bar{\Lambda}^t$ on the space of Hermitian 2×2 matrices with zero trace (cf. $\mathrm{SU}(2) \cong \mathrm{Spin}(3)$). Therefore, the general regular tetrahedron may be obtained by allowing such a Λ to act on the matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} 1 & i \\ -i & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1+i \\ 1-i & 0 \end{pmatrix}$$

and then picking out the top right hand entries. We obtain

$$\Lambda \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} a\bar{a} - a\bar{b} - b\bar{a} - b\bar{b} & 2ab + a^2 - b^2 \\ 2\bar{a}\bar{b} + \bar{a}^2 - \bar{b}^2 & b\bar{b} + a\bar{b} + b\bar{a} - a\bar{a} \end{pmatrix}$$

$$\Lambda \begin{pmatrix} 1 & i \\ -i & -1 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} a\bar{a} - ia\bar{b} + ib\bar{a} - b\bar{b} & 2ab + ia^2 + ib^2 \\ 2\bar{a}\bar{b} - i\bar{a}^2 - i\bar{b}^2 & b\bar{b} + ia\bar{b} - ib\bar{a} - a\bar{a} \end{pmatrix}$$

$$\Lambda \begin{pmatrix} 0 & 1+i \\ 1-i & 0 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} (i-1)b\bar{a} - (i+1)a\bar{b} & (i+1)a^2 + (i-1)b^2 \\ (1-i)\bar{a}^2 - (1+i)\bar{b}^2 & (1+i)a\bar{b} + (1-i)b\bar{a} \end{pmatrix}$$

so

$$\alpha = 2ab + a^2 - b^2, \quad \beta = 2ab + ia^2 + ib^2, \quad \gamma = (i+1)a^2 + (i-1)b^2$$

and, again, it is easy to check that (2) is satisfied. Conversely (2) is exactly the condition that α, β, γ may be written in this way for some $a, b \in \mathbf{C}$ (since every non-singular conic in $\mathbf{CP}_2$ is rational). □

There are several variations on this theme. Each simplex in $\mathbf{R}^3$ gives rise to a quadratic equation which characterises the orthographic images of similar simplices. For example, if a vertex of a cube is mapped to the origin and its neighbours are sent to $\alpha, \beta, \gamma \in \mathbf{C}$, then $\alpha^2 + \beta^2 + \gamma^2 = 0$ (this is due to Hadwiger [H], though not stated in this form using complex numbers). The corresponding equation for a regular dodecahedron is

$$(\alpha + \beta + \gamma)^2 + (\sqrt{5} - 1)(\alpha^2 + \beta^2 + \gamma^2) = 0.$$

There is a general theory for orthogonal projection $\mathbf{R}^n \to \mathbf{R}^m$ with an interpretation in terms of complex numbers when $m = 2$. Details will appear elsewhere.

Thanks are due to H.S.M. Coxeter for drawing our attention to Hadwiger's article and to Jane Pitman for useful conversations.

References

[H] H. Hadwiger, Über ausgezeichnete Vectorsterne und reguläre Polytope, Comm. Math. Helv. **13** (1940), 90–108.

225

AUTHOR'S RESPONSE – AN EMPEROR STILL WITHOUT MIND

Behavioural and Brain Sciences **16** (1993), no. 3, 616–22

There were five of the original reviewers of my book *The Emperor's New Mind* (Oxford University Press, 1989; see Chapter 195), who come back with further commentaries on my original responses (Chapter 196). These are my further responses.

Author's Response

An emperor still without mind

Roger Penrose
Mathematical Institute, University of Oxford, Oxford OX1 3LB, England

It is a refreshing change to be able to address a mere five peer commentaries, rather than the 37 original ones that were concerned with issues raised in my book *The emperor's new mind: Concerning computers, minds, and the laws of physics* (Penrose 1989) – hereafter referred to as *Emperor* – and the subsequent *BBS* Précis (Penrose 1990a). The present five commentaries pertain also to this book and to that previous discussion, in particular the Author's Response: "The nonalgorithmic mind" (Penrose 1990r) – hereafter abbreviated: *Emperor Response*. I welcome this opportunity to comment on them, and I hope that I am able to give a better and more constructive response to these five than I was able to before, for the earlier 37. I have no doubt that the sheer magnitude of the earlier commentary resulted in my remarks, at that time, being all too brief and frequently inadequate, with even some significant points having been left unaddressed. May I take this opportunity to apologize to those commentators who were (or who may well have felt themselves to be) given short shrift by my earlier remarks.

The present commentaries are by **Davis, Guccione, Squires, Wallace,** and **Wilson,** taking them in alphabetical order. I shall address these commentaries essentially in turn, in that order.

The commentary by **Davis** is very helpful in that it brings out several points that certainly need further discussion. I believe that there are various misunderstandings between us, rather more than actual disagreements, and I shall do my best, in the following remarks, to try to sort these out. In fact, **Guccione**'s commentary constitutes a very useful reply to the main point brought up by Davis, so it will be helpful if I may give my response to Guccione in conjunction with that to Davis.

I hope that **Davis** did not take my earlier response about the multifarious misunderstandings of what I am calling the "Gödelian" argument as being explicitly addressed to the working logicians among my critics: Boolos (1990) and himself Davis (1990). It was hard for me to make distinctions when I had so many different people to reply to. It was clear, for example, that some of my critics had not even grasped the import of the *reductio ad absurdum* nature of the Gödelian argument as given, and there were other misunderstandings at a similar level. I was certainly not trying to suggest that Boolos or Davis could have misunderstood what I was trying to say at this kind of level. Nevertheless it seems clear that there *are* still misunderstandings between Davis and me, as his present comments would appear to indicate.

Let me begin by thanking **Davis** for supplying the correct logician's terminology, namely Π_1-*sentences*, for what I had referred to as "propositions of the form $[Q]$." I am happy to adopt his terminology henceforth. The terminology I had used when referring to the "Gödelian *insight*," however, was chosen deliberately. I think it would miss the point of what I was trying to express – no doubt very inadequately – by just referring to it as an "easy theorem" within some relatively elementary formal system such as Peano arithmetic. I am prepared to accept that it was misleading for me to try to denote this insight in the partially symbolic form

$$\text{“}F \text{ sound''} \Rightarrow \text{“}G(F) \text{ true,''}$$

as Davis's remarks make clear – and I am very grateful to him for pointing out the result within Peano arithmetic that indicates how my descriptions can indeed mislead. But the insight I intended to express, which comes from Gödel's theorem, appears to get somewhat lost if one simply phrases it as the formal theorem that Davis refers to, because the "insight" involves not just a piece of *formal* mathematics – that is, a succession of strings of symbols correctly obeying the rules of some formal system – but also the *understanding* of the underlying *meanings* that the symbols of the formalism encode (such as the fact that the actual meaning of the symbol "$\forall$" is to be "for all natural numbers," etc.). I think that **Guccione** makes this point very thoroughly and forcefully, and I am grateful to him for his clarifications. I can hardly do better than refer Davis – and anyone else who is adequately versed in the logical notation and background discussion – to Guccione's excellent commentary.

There are, however, several further clarifications that it may be worthwhile for me to attempt to make here. I am sure most of what I shall say is familiar to **Davis**, but I hope these remarks will be helpful in sorting out our remaining differences and will also be helpful to others. As I see it, the theorem Davis refers to would express just one part of what I mean by the "Gödelian insight," without coming to grips with its full meaning. The relevant implication of the Gödelian insight – when "F sound $\Rightarrow G(F)$ true" is formalized within, say, Peano arithmetic, where the Peano arithmetic is being taken to lie within the Zermelo-Fraenkel system (ZF) whose own Gödel sentence is being examined – lies outside the compass of ZF itself. This is essentially because ZF is incapable of asserting its own soundness. A "ZF-person" would say, roughly: "If I am sound, then $G(\mathrm{ZF})$ is true – but since I don't know whether or not I *am* sound (or even what soundness means), I cannot assert that $G(\mathrm{ZF})$ is actually 'true,' in the only sense of the word I know, that is, provable by me.'" Of course, looking at this from the outside, it would be something of an absurdity if it were an actual personal statement, because if one does not know that one is sound, then one cannot trust *any* of one's assertions – even that which asserts that *if* ZF is sound then $G(\mathrm{ZF})$ is true (although I appreciate that this particular implication is derived using only the more reliable "Peano arithmetic" part of ZF). The trust in $G(\mathrm{ZF})$ is, indeed, no less than the trust in other assertions that ZF can make, since these depend upon the assumption of ZF's soundness in any case. Thus, from the outside, we see that even though the "ZF-person" can not actually *assert* $G(\mathrm{ZF})$, it asserts something that in terms of actual *meanings* is equivalent to its asserting $G(\mathrm{ZF})$. If it understood "meanings" then it *could* just as well have asserted $G(\mathrm{ZF})$! The basic problem is that a formal system, unlike a human mathematician, cannot "know" what soundness (or mathematical truth of, say, Π_1-sentences) actually means. (There is an absurdity that is perhaps even more blatant, if we replace "sound" in the above displayed statement by "consistent" – where we take the version of $G[F]$ that actually asserts F's consistency. For if ZF is *not* consistent, then it *can* assert $G[\mathrm{ZF}]$, whereas if it *is* consistent it *cannot*, although then the meaning of its assertion "if ZF is consistent then $G[\mathrm{ZF}]$ is true" implies the actual truth of $G[\mathrm{ZF}]$.

As just mentioned above, these absurdities arise because the actual concept of mathematical *truth* cannot be captured by *any* formal system, as **Guccione** points out (Tarski's result). In order to pass from a belief in F to a belief in $G(F)$, we need to know the *meanings* that are involved. This is something that we supply from the outside, but as far as a formal system itself is concerned, there is no "legal" way, within F, of actually passing from F to $G(F)$. Another way of expressing the point is to say that "beliefs" are not formalizable; for we *can* get from "belief in F" to "belief in $G(F)$," using the intuitive notion of what we mean by "belief," but we cannot *formally* do this within F.

I am not sure, however, whether **Davis** is asserting that the "simple theorem" he refers to is supposed to achieve the actual passage (of one's belief?) from F to $G(F)$. It is certainly true that the procedure whereby one passes from F to $G(F)$ is an algorithmic matter, so one could write a computer program G that, for each F, yields $G(F)$ if it is presented with F. This point has been made many times before, but it in no way makes the "Gödelian insight" into an algorithmic matter – and the fact that it does not is basically why I am calling it an "insight" rather than a "procedure" or even a "simple theorem." The "insight" concerns one's unassailable mathematical beliefs, and the fact that (unassailable) *belief* in F entails (unassailable) *belief* in $G(F)$. What I am saying is that mathematical belief cannot be a known algorithmic matter, by which I here mean that there cannot be an algorithm A that one knows generates precisely one's believed mathematical statements – and, as Davis concurs, let us stick to statements which have the form of Π_1-sentences (i.e., $[Q]$-type). For one's beliefs must include a belief in A itself, and also (by the Gödelian insight) in the *validity* of the particular procedure G, that produces the Π_1-sentence $G(F)$ when presented with F. (If one believes in the soundness of F, then one must believe that $G[F]$ is true.) Thus, if there were such an A, then $G(A)$ would also have to be believed. But $G(A)$ is also believed to lie outside the scope of A (and actually *does* lie outside the scope of A, if A is in fact sound) – by another part of the Gödelian insight – so $G(A)$ cannot be believed after all, which is a contradiction.

Note that, somewhat remarkably, it does *not* need to be the case that the putative algorithm A that is supposed to be underlying one's belief system is actually *sound* for this argument to work. The important thing is just that A's role must be *known:* so that this knowledge forces one to include a belief in the soundness of A amongst one's actual beliefs. Note, also, that the fact that G is an algorithm plays no part in the discussion. (Nor, in fact, need A necessarily be an algorithm either, for this general kind of argument to work; certain specific types of known "oracle" would also do – but that's another story.) The algorithmic nature of the passage from F to $G(F)$ is neither here nor there. The fact that this passage is part of one's *belief* system but not part of F is, however, what puts overwhelming constraints on the knowable nature of such a belief system. These constraints may tend to fade from view if we try to allow the "slippery" Gödelian insight merely to be a simple theorem of Peano arithmetic – for then I fear he may *already* have slipped away!

If it is not labouring the point excessively, let me put things a little differently. Imagine a computer-controlled robot that is correctly able to ascertain the truth of certain Π_1-sentences by means of some sound algorithmic procedure A. To make things more graphic, I shall use an anthropomorphic terminology and say that the robot "knows" the truth of precisely those Π_1-sentences that it can derive by the use of A. It will *not* "know" the Π_1-sentence $G(A)$, however, even though the procedure for obtaining $G(F)$ from F is perfectly algorithmic. Of course, we could *tell* our robot that $G(A)$ is indeed true (using our own insights to that effect), but if the robot were to accept this fact, it would have to modify its own rules by adjoining this new truth (an instance of *our* Gödelian insight) to the ones that it already "knows." We could imagine going further than this and tell our robot, in some appropriate way, that the general algorithmic procedure for obtaining $G(F)$ from F is also something that it should "know" as a way of obtaining new truths from old. Anything that is well-defined and calculational – in particular the algorithm G – could be added to the robot's store of "knowledge." But now the robot operates according to a *new* "A," and the Gödelian insight would apply to this, instead of to the old A. That is to say, we should have been using this new "A" all along instead of the old A, since it is cheating to change our "A" in the middle of the argument. This is the very nature of *reductio ad absurdum*. Whatever A we settle on, its $G(A)$ will be outside the scope of the robot's "knowledge."

The trouble for our poor robot is that in the absence of any *understanding* of the Gödel procedure – that is, not possessing an *actual* "Gödelian insight" – the robot has no reliable independent way of judging the truth of $G(A)$, other than being told by us. This is a separate matter from the calculational aspects of the Gödel argument. To be able to do more than this, it, like us, would need to understand the *meanings* of the operations it has been told to perform. Without understanding, it could equally well (wrongly) "know" that $G(A)$ is *false* instead of true. It is just as much an algorithmic matter to derive (wrongly) that "$G(A)$ is false" as it is to derive (correctly) that "$G(A)$ is true." Thus, the algorithmic nature of these operations is not the point; the point is that our robot requires sound "*beliefs*" in order to know which algorithms give it truths and not falsehoods. These "beliefs" could include *some* instances of the Gödelian insight, but *not* that particular instance that would allow it to "believe" $G(A)$ even though it ought, in the ordinary sense of the word, to "believe" its own belief system. Thus it cannot encompass the *entire* Gödelian insight.

Davis refers to *iteration* of the Gödelization procedure at the beginning of his commentary. Although I discussed this procedure in *Emperor* (pp. 109–10) and gave reference both to Turing (1939/1965) and to Feferman (1988), I did *not* use it there to support my contention that Gödel's theorem provides a powerful case in favour of the non-algorithmic nature of mathematical insight. On the other hand, this iteration procedure *does* illustrate something of the fact that the "Gödelian insight," in the sense that I mean it, *cannot* be algorithmic. For Feferman showed, if I understand him rightly, that given any true Π_1-sentence S, there is an iterated Gödelization starting from, say, Peano arithmetic, that yields a sound formal system in which S is a theorem. If the general procedure of iterated Gödelization were algorithmic, then there would be an algorithm for generating the true Π_1-sentences. But there is no such algorithm; hence iterated Gödelization is not algorithmic. The trouble comes, as I know Davis is well aware, from the fact that there is no algorithmic way of generating the recursive ordinals, these ordinals being the objects that characterize the different degrees of iterated Gödelization. In each instance, one would have to perceive exactly how to incorporate the particular ordinal in question into an iterated Gödelization procedure, but there is no uniform algorithmic way of doing this. This is the general kind of thing that I had in mind in *Emperor Response*, p. 695, when I referred to:

> The slipperiness of the nonalgorithmic nature of the Gödelian insight lies partly in the fact that we are never quite sure what it is that has slipped through the algorithmic net, though something always does! One might argue that what has slipped through it is the very perception of how one actually formulates something in an algorithmic way.

This perception, in the case of iterated Gödelization, is of how one actually formulates the algorithmic procedure of the iteration, for any given recursive ordinal. Since there is no *general* algorithm for this, one's mathematical insights need to be brought in, illustrating again that mathematical insight is, in general, not an algorithmic matter. What the Gödelian insight allows one to do is to show how one can leap to a new set of procedures once the algorithmic scope of one's previous insights has been fully incorporated into one's previous procedures.

We now come to the question of my slapdash scholarship. I think that if **Davis** looks at *Emperor* carefully (p. 34; also footnote on p. 129) he will find that I did not actually conflate Hilbert's 10th problem with the *Entscheidungsproblem*, though it is easy to misread my account in this way – especially as I may not have given the most appropriate reference to the latter. (I apologize for not having given credit to Martin Davis, Julia Robinson, and Yuri Matyasevich, on p. 129, for their wonderful resolution of Hilbert's 10th problem [Davis et al. 1976], especially because hasty readers might perhaps misread my account as suggesting that Turing had actually resolved Hilbert's 10th problem!) Let me now address the question of what it was that I was claiming, in *Emperor Response* (p. 693), that Gödel "certainly appreciated . . . in the 1930s." I think there is another genuine misunderstanding here, and since a number of other people have also misunderstood me in the same way as Davis, I should undoubtedly try to set the record straight as to what I really intended.

What I said in *Emperor Response* was in reply to the multitude of claims that in my use of the Gödel argument I was simply trotting out "old" fallacious arguments. Most of these claims were not addressing subtle issues of interpretation, but were simply based on misunderstandings of my use of the Gödelian argument (such as its *reductio ad absurdum* nature, as mentioned above). My actual words (Response, p. 693), in reference to my own arguments, were: ". . . I make no apology for the fact that much of it is 'old,' certainly appreciated by Gödel himself already in the 1930s and never properly refuted since." What I was basically referring to here was the *part* of the argument (". . . much of it . . .") that I had taken to be an unexceptionable application of Gödel's incompleteness theorem, providing us with the conclusion, as given in *Emperor Response*, p. 696: ". . . humans do not ascertain mathematical truth by means of any knowable algorithm"; and, a little lower down, expressed somewhat more carefully: "The algorithm . . . X [according to which mathematicians as a whole are supposed to act], or at least the soundness . . . of X, would have to be something unknowable to human mathematicians. Otherwise $G(X)$ would be seen to be a mathematical truth, although it is inaccessible to *X*: a contradiction." In *Emperor* itself, p. 418, my conclusion from Gödel's theorem is: "Thus, we are driven to the conclusion that the [putative] algorithm that mathematicians actually use to decide mathematical truth is so complicated or obscure that its very validity can never be known to us."

Far from being contradicted by what Rucker (1982) quotes Gödel (1988/1990) as saying (as given by **Davis** in his commentary), it seems to me that what Gödel asserts is completely in line with my own quotes. One must bear in mind that Gödel – whose *opinion* was certainly that the mind was not constrained by algorithms – is rather despairing of the fact that his incompleteness theorems do not fully prove this point of view, and he is admitting that "on the basis of what has been proved so far," an algorithm such as my *X* ("a theorem-proving machine") could exist, "but cannot be *proved* to be" "equivalent to mathematical intuition." We must assume that Gödel was taking his mathematical argument as far as he could, so he must certainly have been aware that there could be no algorithm that could be *known* to us to be equivalent to mathematical intuition, which is virtually identical to my own claims. In fact, Gödel uses "proved" where I have used "known" (or, in this account, "believed" or "unassailably believed") but the meanings are essentially the same. I have preferred to use "known" rather than "proved" because I was trying to steer clear of the connotation that "proved" often refers to arguments given within some specified formal system. (Also, I use "obscure" where others might use "nonconstructive," or something of this nature.) But I do not think there is any great difference between our viewpoints here as to the philosophical implications of the Gödel theorems.

This was all that I meant was "certainly appreciated by Gödel," but even this much was certainly *not* appreciated by many of my *BBS* critics last time, to whom my remarks were directed. In fact, even for this much of the argument there are a good many possible loopholes that need to be addressed. (Addressing them all takes more than 30 pages in a new book that I am presently engaged in writing – henceforth referred to as *Shadows*.) In proceeding beyond this, the arguments admittedly get more speculative, and in *Emperor Response*, p. 696, I say: "I never claimed that the existence of *X* is a mathematical impossibility, but it would seem to be exceedingly unlikely" and the rest of the argument is of the nature of a strong plausibility argument, aimed at demonstrating that *X*, if not logically impossible, is at least very hard to believe in. Thus, I certainly do not agree with **Davis** when he says that Gödel's theorem has nothing to say about the algorithmic nature of mathematical insight. It has a *great deal* to say about it, although the theorem needs a good body of further supporting argument (which I shall also try to provide in *Shadows*). Davis and I are certainly in agreement that it is an important and difficult question.

Two final points should be made in connection with **Davis**'s commentary. They are both in the nature of apologies. The first apology is for misreading him as being "of a formalist disposition" – and I am glad that he, too, regards Platonism as a more sensible mathematical standpoint. I do not think, however, that I was taking the Platonist position as being inconsistent with a mechanist view of mind.

My second apology is that I did not address, in *Emperor Response*, the suggestion made by Davis (1977) that the ideas of the logician Takeuti (1978) might be able to resolve some of the puzzles ("magic") involved in the replacing of classical variables by operators when one passes from classical to quantum mechanics. I agree with **Davis** that one should be on the lookout for ways in which ideas from mathematical logic may find significant application in physics, and that we may reasonably expect the ultimate impact of mathematical logic upon basic physical theory may indeed be profound. It is fair to say, however, that not a great deal along these lines has come to light as yet, although there have been interesting ideas put forward from time to time (cf. Komar 1964, for example, in which an attempt is made to relate quantum field theory to noncomputability). The suggestion Davis makes is also an interesting one, but it seems to me that its applicability, as things stand, is somewhat limited. As I understand the situation, the Takeuti-Davis proposal only gives clearly unambiguous results when the operators commute, which is a very limiting requirement. Noncommuting operators are the rule in quantum theory, and there are deep issues relating to the uniqueness, or otherwise, of the very process of quantization. The actual *choice* of variables to be quantized, for example, can make a considerable difference to the quantum theory that is obtained (cf. Woodhouse 1992). It is unclear to me that the Takeuti-Davis proposal could help with these basic issues, but it would certainly be of great interest if it can.

Now let me turn to the remaining commentaries. I must apologize to **Squires, Wallace,** and **Wilson** for devoting so little space to them, in comparison with my **Davis/Guccione** response. Indeed, all that I can say about the commentary of Squires is that I wholeheartedly agree with all that he says. He makes a very pertinent remark when pointing out that those who claim to maintain a *"physicalist"* philosophical position rarely have much grasp of the basic facts and the profound problems of the present-day quantum-mechanical description of what we understand to be *physical* reality. He finds it hard to see how so many philosophically minded people seem to believe they can hope to understand the mind while totally ignoring such questions – and I agree! His contribution is especially welcome, for he writes with the authority of a physicist who has devoted a lifetime to the study of the implications of quantum theory. The fundamental importance of quantum theory in our understanding of what "physicality" actually is – and the equally fundamental limitations of that very theory, as we understand it today – cannot really be ignored by anyone who claims a physicalist stance.

Both **Wallace** and **Wilson** are, like **Squires,** concerned with quantum theory, though in their case they are more directly addressing the puzzling issue of how quantum effects might actually be relevant in brain action. Wallace makes a number of suggestions in the positive direction of trying to see how quantum effects might be made use of in the brain, whilst Wilson's comments are more negatively oriented, in that he points out severe problems confronting certain such ideas.

Wallace's suggestions look intriguing, although it is beyond my competence to assess a good part of what he writes. On the other hand, when he considers that certain structures in the brain might act as *quantum computers*, I fear he underestimates the very serious difficulties in constructing such objects and overestimates their presently understood theoretical capabilities.

For a quantum computer to work, it would have to be very strongly isolated from its surroundings, so that the several computations it performs simultaneously in quantum linear superposition do not get inappropriately coupled to those surroundings, for if that happens, the state gets "reduced" – by what I call process **R**, in *Emperor* – and only one of the superposed alternatives survives. It is very difficult, in general, to preserve quantum superpositions except between states that are very similar to one another; and once the differences between even these are "noticed" by the outside world, reduction **R** takes place. With a quantum computer, we must not require that the superposed states be completely similar to each other, because they are supposed to represent the carrying out of independent calculations.

The other difficulty concerns what a quantum computer is theoretically capable of doing. Whereas it is true that it acts as a single device carrying out many calculations simultaneously in superposition, the difficulty comes when one tries to extract the information of these calculations from the device. The extraction of the information would consist of performing a "measurement" on the device, which means allowing the process **R** to take place in an appropriately controlled way. All one could extract would be the answer to a single carefully organized *combination* of the answers to the various calculations; one cannot extract the results of all the calculations individually. Thus, if one imagined that the device, in attempting to solve an instance of the "travelling salesman" problem, would simultaneously try huge numbers of possibilities, one would somehow have to devise a single measurement on the device that would extract merely the appropriate minimizing solution out of the superposition of individual answers. No way of doing such a thing is in evidence. In his original rather artificial example of a problem in which a quantum computer could, in a certain rather marginal sense, achieve a better performance than a classical computer, Deutsch (1985; 1986; cf. also Jozsa 1991) needed to resort to a somewhat artificial-looking problem, rather than something clearcut like the travelling salesman problem, for just such a reason as this. In a more recent paper, Deutsch and Jozsa (1992) have improved upon Deutsch's original example, and here the quantum computer *does* provide a significant improvement over any classical one, in the complexity-theoretic sense. The problem is still a rather special and artificial-looking one, however, and it is not yet known whether a quantum computer could outperform a classical one in solving an NP-complete problem like the travelling salesman one.

A further but separate point is that quantum computers, as the term is technically understood, cannot exceed the bounds of *computability* theory (Deutsch 1985), even though they seem to have some advantages with respect to complexity theory. Though the latter is interesting, and could well eventually become very important (as **Wallace** suggests), from my own point of view, we need to find ways in which the brain is able to do noncomputable things also. (See my comments in relation to **Davis/Guccione** above.) This is why I believe we need to look not just to quantum effects in brain action but to effects that make use of whatever it is that is *really* going on when in present-day quantum theory we use the **R** procedure. With a quantum computer, one envisages a system that behaves entirely as a quantum system, evolving according to **U** (the Schrödinger equation) right up to the time when the final measurement is made, at which point the random action of **R** comes into play. What I believe we *really* need is something that goes *beyond* a quantum computer, in which some noncomputable and not entirely random action comes in at the **U**/**R**-borderline level. This would require some new basic physical theory to understand what is going on, but, as **Squires** has emphasized, some basically new physical theory is indeed needed at some such level in any case. (My own view is that this new theory must be an integral part of quantum gravity, but there is no need to be tied to any particular viewpoint at this stage of understanding.) In the absence of such a new theory, however, it is certainly significant to examine how the existing quantum-computer idea might have importance in brain action, as Wallace is at least trying to do. I wish him luck!

Wilson, though also strongly sympathetic to the idea that one must turn to quantum theory if one is to have any hope of comprehending the phenomenon of consciousness in scientific terms, is very appropriately cautionary in pointing out difficulties confronting various suggestions I had hinted at in *Emperor*. Although he stresses the importance of the global (or "holistic") features of quantum theory in relation to the problem of the "oneness" of conscious introspection, his detailed discussion seems more concerned with the problem of how an individual quantum event (that we might choose to picture as being "influenced," in some way, by consciousness) could affect the larger-scale neuronal activity of the brain. Perhaps my account in *Emperor* was somewhat misleading when I stressed the fact that there *are* neurons – actually present in the retina – that can be triggered by single quantum events. I was trying to point out that the role of quantum theory cannot be ignored in the study of neurophysiological activity, but I think that already by the time I came to write that part of *Emperor* I had formed the view that single-quantum enaction of neuronal activity was, in itself, unlikely to lead us toward answers to the puzzles of consciousness. Nevertheless, there remained the possibility that neurons of this kind could still have some importance for these questions, so their presence, or otherwise, in parts of the brain proper would be a matter of considerable interest. Wilson's points are well taken here, and I am inclined to agree with him that we should not look for an "influence of the mind" on individual quantum events in order to explain the overall behaviour of the brain.

The problem, however, is that at present we have very little indication of an essentially quantum-mechanical process in brain function – even though many of us believe it must be there somewhere. As **Wilson** concurs, although there is much about the underlying physics of the brain – such as even the very stability of atoms – that, strictly speaking, *does* need quantum mechanics for its explanation, this is not the kind of essentially quantum-mechanical role we are looking for. In *Emperor*, p. 438, I speculated that we would need something in which "not just one of the alternative arrangements is tried out, but vast numbers, all superposed in complex linear superposition. . . . If kept below this [one-graviton] level, simultaneous superposed calculations can be performed, very much in accordance with the principles of a quantum computer." The thing that most troubled me about this (and still does) is the difficulty in keeping such superpositions isolated from their surroundings, as I have already remarked in my response to **Wallace** above. In the terms of the aforementioned quotation from *Emperor*, the difficulty would be in keeping the total disturbance "below the [one-graviton] level."

The trouble in this respect with nerve signals is that they seem to be much too macroscopic – that is, they disturb their surroundings much too much – for such superpositions to be maintained. My own viewpoint has undergone a significant shift in recent months regarding the most likely place to look for sufficiently isolated quantum-superposed "calculations." Having been influenced by the researches of Hameroff (1987) and his coworkers, I am impressed by the apparent importance of the cytoskeleton in controlling and maintaining the strengths of synaptic connections. In one-celled animals such as paramesia it is the controlling activity of the cell's cytoskeleton that enables the creature to behave in very complex ways, and apparently also to learn from experience, even though it possesses no nervous system at all. In an individual human neuron, cytoskeletal fibres span its entire length and it strikes me as at least distinctly *possible* that the necessarily isolated quantum coherence, involving some kind of complicated computational(?) behaviour, might be taking place within such fibres. In view of this necessary isolation, however, it would *in principle* be very hard to detect such activity from the outside – so, according to this view, conscious thoughts would be very much private affairs! I should emphasize that I am not looking to tiny quantum influences of the "mind" on individual cytoskeletal fibres, but to a "holistic" type of quantum coherence between very large numbers of such fibres that would, in some appropriate sense, actually "constitute" the mind. I intend to discuss these matters in much greater detail in *Shadows*, and I hope it will then become clear to **Wallace** that although Platonism has an important role to play in my thinking this role is in no way that of a *deus ex machina!*

ACKNOWLEDGMENT
I am grateful to Richard Jozsa and Angus MacIntyre for useful comments.

References

Albert, D. & Loewer, B. (1988) Interpreting the many-worlds interpretation. *Synthèse* 77:195–213. [EJS]
Aspect, A. & Grangier, B. (1990) Wave-particle duality: A case study. In: *Sixty-two years of uncertainty,* ed. A. I. Miller. Plenum. [EJS]
Bhaumik, D., Bhaumik, K. & Dutta-Roy, B. (1976) A microscopic approach to the Frohlich model of Bose condensation of phonons in biological systems. *Physics Letters* 59A:77–80. [RW]
Blum, A. & Rivest, R. (1989) Training a 3-node neural network is NP-complete. In: *Advances in neural information processing systems 1,* ed. D. Touretzsky. Kaufmann. [RW]
Bohm, D. J. & Hiley, B. J. (1984) The de Broglie pilot wave theory and further development of new insights arising out of it. In: *Quantum, space and time – the quest continues,* ed. A. Barut, A. van der Merwe & J. P. Vigier. Cambridge University Press. [EJS]
Bond, J. & Huth, G. (1986) Electrostatic modulation of electromagnetically induced nonthermal responses in biological membranes. In: *Modern bioelectrochemistry,* ed. F. Gutman & H. Keyzer. Plenum. [RW]
Boolos, G. (1990) On seeing the truth of the Gödel sentence. *Behavioral and Brain Sciences* 13(4):655–56. [rRP, MD]
Bower, J. & Llinas, R. (1983) Simultaneous sampling of the responses of multiple, closely adjacent, Purkinje cells responding to climbing fiber activation. *Society for Neuroscience Abstracts* 9:607. [RW]
Butterfield, J. (1990) Lucas revived? An undefended flank, *Behavioral and Brain Sciences* 13(4):658. [SG]
Davis, M. (1958/1982) *Computability and unsolvability.* McGraw-Hill. Reprinted with an appendix, Dover. [MD]
(1977) A relativity principle in quantum mechanics. *International Journal of Theoretical Physics* 16:867–74. [rRP, MD]
(1990) Is mathematical insight algorithmic? *Behavioral and Brain Sciences* 13:659–60. [rRP, MD]
Davis, M., Matijasevic, Y. & Robinson, J. (1976) Hilbert's tenth problem: Diophantine equations; positive aspects of a negative solution. *Proceedings of Symposia in Pure Mathematics* 28:323–78. [rRP]
Dawkins, R. (1986) *The blind watchmaker: Why the evidence of evolution reveals a universe without design.* Norton. [RW]
Deutsch, D. (1985) Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London* A400:97–117. [rRP, RW]
(1986) In *Quantum concepts of space and time,* ed. R. Penrose & C. J. Isham. Oxford University Press. [rRP]
Deutsch, D. & Jozsa, R. (1992) *Proceedings of the Royal Society* (London) A439:553–58. [rRP]
Eccles, J. C. (1970) *Facing reality.* Springer-Verlag. [DLW]
Feferman S. (1988) Turing in the land of O(Z). *The universal Turing machine – a half-century survey,* ed. Rolf Herken. Verlag Kemmerer & Unverzagt/Oxford University Press. [rRP, MD]
Frohlich, H. (1986) Coherent excitations in active biological systems. In: *Modern bioelectrochemistry,* ed. F. Gutman & H. Keyzer. Plenum. [RW]
Garey, M. & Johnson, D. (1979) *Computers and intractability: A guide to the theory of NP-completeness.* Freeman. [RW]
Ghirardi, G. C., Rimini, A. & Webber, T. (1986) Unified dynamics for microscopic and macroscopic systems. *Physical Review* D34:471–91. [EJS]
Gödel, K. (1986/1990) *Kurt Gödel, collected works,* vol. 1 (orig. pub. 1929–1936) & vol. 2 (orig. pub. 1938–1974), ed. S. Feferman, J. W. Dawson, Jr., S. C. Kleene, G. H. Moore, R. M. Solovay & J. van Heijenoort. Oxford University Press. [rRP, MD]
Gray, J. A. (1982) Précis of *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. Behavioral and Brain Sciences,* 5:469–534. [RW]
Hameroff, S. (1987) *Ultimate computing: Biomolecular consciousness and nanotechnology.* North Holland. [rRP]
Higginbotham, J. (1990) Penrose's Platonism. *Behavioral and Brain Sciences* 13(4):667–68. [SG]
Jozsa, R. (1991) Characterizing classes of functions computable by quantum parallelism. *Proceedings of the Royal Society (London)* A435:563–74. [rRP]
Kaiser, F. (1978) Coherent oscillations in biological systems II. Limit cycle collapse and the onset of travelling waves in Frohlich's brain wave model. *Zeitschrift für Naturforschung* 33a:418–31. [RW]
Kandel, E. (1976) *Cellular basis of behavior.* W. H. Freeman. [RW]
Kentridge, R. W. (1990) Parallelism and patterns of thought. *Behavioral and Brain Sciences* 13(4):670–71. [SG]
Komar, A. B. (1964) Undecidability of macroscopically distinguishable states in quantum field theory, *Physical Reviews* 133B:542–44. [rRP]
Lashley, K. S., Chow, K. L. & Semmes, J. (1951) An examination of the electrical field theory of cerebral integration. *Psychological Review* 58:123–36. [DLW]
Libet, B., Gleason, C. A., Wright, E. W., Jr. & Pearl, D. K. (1983) Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential); the unconscious initiation of a freely voluntary act. *Brain* 106:623–42. [DLW]
Lockwood, M. (1989) *Mind, brain, and the quantum: The compound I.* Basil Blackwell. [RW]
MacVicar, B. & Dudek, F. (1980) Dye-coupling between CA3 pyramidal cells in slices of rat hippocampus. *Brain Research* 196:494–97. [RW]
Marshall, I. (1989) Consciousness and Bose-Einstein condensates. *New Ideas in Psychology* 7(1):73–83. [RW]
McNaughton, B., Leonard, B. & Chen, L. (1989) Cortico-hippocampal interactions and cognitive mapping: A hypothesis based on reintegration of the parietal and inferotemporal pathways for visual processing. *Psychobiology* 17:230–35. [RW]

Miller, R. (1989) Cortico-hippocampal interplay: Self-organizing phase-locked loops for indexing memory. *Psychobiology* 17:115–28. [RW]

Montague, R. (1962) Theories incomparable with respect to relative interpretability. *Journal of Symbolic Logic* 27:195–211. [SG]

O'Keefe, J. (1985) Is consciousness the gateway to the hippocampal cognitive map? A speculative essay on the neural basis of mind. In: *Brain and mind*, ed. D. A. Oakley. Methuen. [RW]

O'Keefe, J. & Nadel, L. (1978) *The hippocampus as a cognitive map.* Oxford University Press. [RW]

Peierls, R. (1991) In defence of measurement. *Physics World* 4:19–20. [EJS]

Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics.* Oxford University Press. [rRP, MD, SG, DLW, RW]

(1990a) Précis of *The emperor's new mind: Concerning computers, minds, and the laws of physics. Behavioral and Brain Sciences* 13:643–55. [rRP, MD, SG, EJS, DLW, RW]

(1990r) The nonalgorithmic mind [Response to commentary]. *Behavioral and Brain Sciences* 13:692–705. [rRP, MD, SG]

(1994) *Shadows of the mind.* Oxford University Press (in press). [rRP]

Potts, R. (1984) Home bases and early hominids. *American Scientist* 72:338–47. [RW]

Putnam, H. (1960) Minds and machines. In: *Dimensions of mind*, ed. S. Hook. New York University Press. [SG]

Rose, S. (1973) *The conscious brain.* Knopf. [RW]

Rucker, R. (1982) *Infinity and the mind: The science and philosophy of the infinite.* Paladin Books/Granada Publishing. [rRP, MD]

Shipman, P. (1986) Scavenging or hunting in early hominids: Theoretical framework and tests. *American Anthropologist*, 88:27–43. [RW]

Smullyan, R. M. (1961) *Theory of formal systems.* Princeton University Press. [SG]

Sperry, R. W. (1958) Physiological plasticity and brain circuit theory. In: *Biological and biochemical bases of behavior*, ed. H. F. Harlow & C. N. Woolsey. University of Wisconsin Press. [DLW]

Squires, E. J. (1987) The unique world of the Everett version of quantum theory. *Foundations of Physics Letters* 1:13–20. [EJS]

(1990) *Conscious mind in the physical world.* Adam-Hilger. [EJS]

(1993) Quantum theory and the relation between conscious mind and the physical world. *Synthèse* 97 (in press). [EJS]

Stapp, H. P. (1990) Quantum measurement and the mind-brain connection. University of California at Berkeley preprint LBL-29594. [EJS]

Takeuti, G. (1978) *Two applications of logic to mathematics.* Iwanami Shoten and Princeton University Press. [rRP, MD]

Teyler, T. & DiScenna, P. (1985) The role of hippocampus in memory: A hypothesis. *Neuroscience and Biobehavioral Reviews* 9:377–89. [RW]

Turing, A. (1939/1965) Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, (ser. 2) 45:161–228. Reprinted: *The undecidable*, ed. M. Davis. Raven Press. [rRP, MD]

Wallace, R. (1989) Cognitive mapping and the origin of language and mind. *Current Anthropology* 3:518–26. [RW]

Witter, M. (1989) Connectivity of the rat hippocampus. In: *The hippocampus: New vistas*, ed. V. E. Chan-Palay. Alan R. Liss. [RW]

Woodhouse, N. M. J. (1992) *Geometric quantization* (2nd. ed.). Oxford University Press. [rRP]

Wu, T. M. & Austin, S. (1977) Bose condensation in biosystems. *Physics Letters* 64A:151–52. [RW]

226
SQUARE-ROOT VOTING

In quest of a better China: Selected essays, discussions and comments from an international conference. Ed. C. Hong-Mo, 86–9. World Scientific, Singapore, 1993

In 1946, my father, Lionel S. Penrose (see Chapter 268), suggested this voting system which is applicable in situations where (as in the United Nations General Assembly) there are blocks of widely differing population sizes, each represented by a single delegate. His suggestion, that each such delegate should have a voting power that is in proportion to the square root of the population size being represented, seems to me to be very fair and it is powerfully supported by precise statistical considerations. I was invited to present these ideas in the somewhat different context of a possible future electoral system for China by my particle-physicist friend and colleague Chan Hong-Mo.

Square-root voting. *In quest of a better China: Selected essays, discussions and comments from an international conference.* Ed. C. Hong-Mo, 86–9. 1993

Square-Root Voting

by Roger Penrose
Mathematical Institute
Oxford, U.K.

Consider a situation where a large population is divided naturally into smaller groups, where the individual groups may differ markedly in size. The groupings may be defined either geographically or according to some other clearly identifiable criterion. I wish to consider the question of how to represent the different groups most fairly in an overall parliament, taking into account possibly large inequalities in the group population size. Let us suppose that each group is represented in parliament by a single delegate whose vote counts with a strength that is dependent upon the size of the population of the group (or else by a number of separate delegates, each with one vote, who are assumed all to vote the same way as each other on every issue). How large should the strength of each delegate's vote be (or, equivalently, how many unanimous but singly-voting delegates should each group provide) according to the size of the group in question?

This issue is of relevance to the United Nations General Assembly, where each member country supplies a single delegate, the individual delegates having votes of equal strength. This would seem to be unfair, since a large country like China has exactly the same voting strength as a small country like Iceland. At a different extreme, we have the situation that arises at the British Labour Party Conference or TUC. Here the various trade unions have block votes of a strength that is proportional to the size of the union. This also seems unfair, but now in the opposite direction, because the large unions appear to exert an excessive control over the decisions that are voted upon. Indeed this impression is correct, since the leadership (or the decisions imposed upon the leadership) of each union are decided upon by a majority vote, rather than by a unanimous vote. If the individual union members were themselves to vote on each issue, instead of delegating their votes to their union leader, then their votes would be divided more evenly, rather than going unanimously (or at least with fixed proportion) in one direction or the other as the block vote system tacitly assumes. Similar remarks apply to the U.S. presidential elections, where individual states have votes (roughly) in proportion to their population size.

An alternative system, which is intermediate between these two extremes, and which has a significant statistical theoretical justification, was put forward over 45 years ago by my father, Lionel S. Penrose (1946, 1952), who was at that time Professor of Human Genetics at University College, London. His proposal was to give each delgate a voting strength that would be in proportion to the square root of the population size that the delegate represents. The statistical justification of this depends upon what he called the "voting power" of a resolute individual in a group that is otherwise assumed (as a not inappropriate approximation) to vote randomly. This power represents the probability that this individual voter's vote will actually affect the result. For a reasonably large group, of size n, this power turns out to be

$$\frac{1}{\sqrt{(n\pi/2)}} = \frac{0.7978}{\sqrt{n}} .$$

Thus, if we want to ensure that the power of each individual's vote is the same, no matter how large is the group that he/she happens to belong to, we must ensure that the voting strength of the group itself is proportional to √n. Not only does this choice have a good theoretical justification, but it also "looks reasonable". For example, using the population sized that were available to my father in 1952, he suggested U.N. voting strengths (rounded up) as follows:

China 22 (pop. 463 million),
India 19 (pop. 339 million),
U.S.S.R 13 (pop 156 million),
U.S.A. 13 (pop. 144 million),
....,
Germany 9 (pop. 67 million),
U.K. 8 (pop. 50 million),
....,
Albania 2 (pop. 1 million),
Iceland 1 (pop. 0.1 million).

A relevant question is whether such ideas might usefully be applied to a voting system within a democratic China. One possibility would be to apply such a system where the different provinces constituted the various "groups". However this would not address the issue that the Han Chinese form easily the dominant ethnic group, and smaller ethnic groups might feel that their own points of view have no chance of being represented in view of such Han domination. Thus, we could try, instead, a system whereby the "groups" under consideration are "ethnic groups". (The situation is perhaps closer to that of the U.K. trade unions, rather than the U.N., where the groups are being distinguished according to a non-geographic criterion.)

Using the 1982 figures (Banister 1987), we find:

Han 305 (pop. 936.7 million)
Zhuang 36 (pop. 13.4 million)
Hui 26 (pop. 7.2 million)
Uygur 24 (pop. 6.0 million)
Yi 23 (pop. 5.5 million)
Miao 22 (pop. 5.0 million)
Manchu 20 (pop. 4.3 million)
Tibetan 19 (pop. 3.9 million)
Mongolian 18 (pop. 3.4 million)
Tujia 16 (pop. 2.8 million)
Bouyei 14 (pop. 2.1 million)
Korean 13 (pop. 1.8 million)
Dong 11 (pop. 1.4 million)
Yao 11 (pop. 1.4 million)
Bai 10 (pop. 1.1 million)
Hani 10 (pop. 1.1 million)
Kazak 9 (pop. 0.9 million)
Dai 9 (pop. 0.8 million)
Li 9 (pop. 0.8 million)
Lisu 6 (pop. 0.5 million)
She 6 (pop. 0.4 million)
Lahu 5 (pop. 0.3 million)
....
Yugur 1 (pop. 0.01 million)
Baoan 0 (pop. 0.009 million)
....,

where, here, I have rounded down rather than up.

I think that one should be a little careful about applying these ideas here, however, for various reasons. In the first place, perhaps people should not be encouraged to classify themselves in one specific ethnic group or another. Ideally, one should be trying to break down such clear-cut divisions. Perhaps there should be another “ethnic classification” that anyone should be entitled to elect to belong to, namely “none”. The hope might be that the “none” category would continue to grow until it became the majority! A second point of worry could the validity of the “randomness” assumption here, in the justification of the square-root rule. Some might feel that ethnic groups might tend to vote more significantly as a body, rather than just randomly. My own standpoint on this particular issue is that the square root system would be good, from this standpoint, since it would tend to beak down this kind of unthinking ethnic unanimity.

Anyway, it’s food for thought.

References

Penrose, L.S. (1946) The Elementary Statistics of Majority Voting J. Roy. Stat. Soc. 109, 53-7.

Penrose, L.S. (1952) On the Objective Study of Crowd Behaviour (H.K.Lewis & Co. Ltd., London).

Feilding, G. and Liebeck, H. (1974) Voting Structures and the Square Root Law B. J. Pol. S. 5, 249-63.

See also letters in The Guardian 23/1/81, 27/1//81, 4/2/81, 12/2/81, 14/2/81, 21/2/81, 10/3/81.

Banister, J (1987) China's Changing Population (Stanford Univ. Press, Stanford, Calif.).

227

ON BELL NON-LOCALITY WITHOUT PROBABILITIES: MORE CURIOUS GEOMETRY

Stud. His. Philos. Sci. **24** (1993), no. 5; *Studies in history and philosophy of modern physics*, 697–720 (1994). With J. Zimba

This paper was actually written (and submitted) considerably later than that which constitutes Chapter 291. The material in the present chapter may be considered to be a follow-up of that other work, and also of Chapter 228 and Section 5.3 of my book *Shadows of the Mind* (Oxford University Press, 1994).

376 On Bell non-locality without probabilities: More curious geometry. *Stud. His. Philos. Sci.* **24** (1993), no. 5; *Studies in history and philosophy of modern physics*, 697–720 (1994). With J. Zimba

On Bell Non-Locality Without Probabilities: More Curious Geometry

Jason Zimba and Roger Penrose†*

1. Introduction

GLEASON'S Theorem (Gleason 1957) implies that in a Hilbert space of dimension $n \geq 3$, it is impossible to assign 1s and 0s to all the rays in such a way that

A. No two orthogonal rays are both assigned the value 1; and

B. In any group of *n* mutually orthogonal rays, not all of the rays are assigned the value 0.

In his famous review paper, Bell (1966) used this fact to prove the inconsistency of non-contextualist deterministic hidden-variable theories. This result, often called the Bell–Kochen–Specker Theorem, was discovered independently at about the same time by Kochen and Specker (1967), who demonstrated the inconsistency in a *finitary* way, explicitly defining a 'non-colourable' set of 117 rays in three-dimensional Hilbert space (where the numbers 1 and 0 are naturally thought of as being two colours, say white and black).

Some years later, by using EPR pairs of such systems, Heywood and Redhead (1983) and Stairs (1983) showed how such a set could be used to prove the inconsistency of *local* deterministic hidden-variable theories.[1] This provided a new (and non-probabilistic) version of Bell's Non-Locality Theorem. Other versions of Bell's Theorem without probabilities have followed more recently (Greenberger, Horne and Zeilinger 1989, Brown and Svetlichny 1990, Clifton 1991). Brown (1991) gives a review of these and other results.

---

* Now at University of California, Berkeley, CA, U.S.A.
† Mathematical Institute, 24–29 St. Giles, Oxford OX1 3LB, U.K.
Received 4 *May* 1993; *in revised form* 25 *June* 1993.

[1] We are informed by A. Shimony that Kochen and Specker had been aware of this fact some time earlier.

Kochen and Specker's example of a non-colourable set of rays generates numerous problems of mathematical interest in finding other examples, especially ones with fewer members or interesting symmetries. Recently, Peres (1991) has given non-colourable sets of 33 rays in three dimensions and 24 rays in four dimensions, and Conway and Kochen (unpublished) found a set of 31 in three dimensions. Penrose (1992) gives a different set of 33 rays in three dimensions, as well as a set of 40 in four dimensions.

When these 40 rays are taken to represent particular spin states for an atom of spin 3/2, only 20 of them (referred to later as the 'explicit' states) turn out to be directly measurable in the normal way by a Stern–Gerlach-type of apparatus. The remaining 20 (the 'implicit' states) do not need to be measured, however, if EPR pairs of such atoms are used (see Penrose 1992 for details). In this article, in common with most of the others cited, we are not concerned with the issue of how one might actually measure the states in question. The Hilbert-space vectors will be treated entirely abstractly, with no reference to the particular type of physical system (e.g. atoms of spin 3/2) that might be used to realize them. Moreover, we shall not be concerned with the possibility of using EPR pairs in order to reduce the number of Hilbert-space directions that actually need to be (potentially) measured.

In this article we begin by briefly describing the 40-ray configuration (Section 2). This set, which has an intriguing geometry based on the dodecahedron, is the first that has been shown to be essentially *complex*—that is, it cannot be expressed solely in terms of vectors with real components.[2] We then show how this configuration can be used to establish Bell's Theorem without probabilities (Section 3). And finally, making use of a complete characterization of the orthogonalities in this set (Section 4), we establish that among the 40 rays there is actually a non-colourable subset, containing 28 rays (Section 5). This configuration, whose geometry is based on the dodecahedron and cube, is also essentially complex; moreover, it is *critical* in the sense that all of its proper subsets are colourable.

At the end of these primary considerations, we include in Section 6 a simple proof of an interesting fact, namely that finite non-colourable sets exist in Hilbert spaces of all finite dimensions $n \geq 3$—except possibly for the $n = 5$ case.[3]

---

[2] Although in this article we are dealing only with the 'dodecahedral' set of 40 rays, for the sake of completeness we remark that Penrose's set of 33 rays in three dimensions is also complex, whereas the other configurations listed can all be given in terms of real vectors.

[3] To avoid making repetitive references in the text, we note here that Sections 2 and 3 are essentially composed of material from Penrose (1992), and that Sections 5 and 6 are essentially composed of material from Zimba (1993).

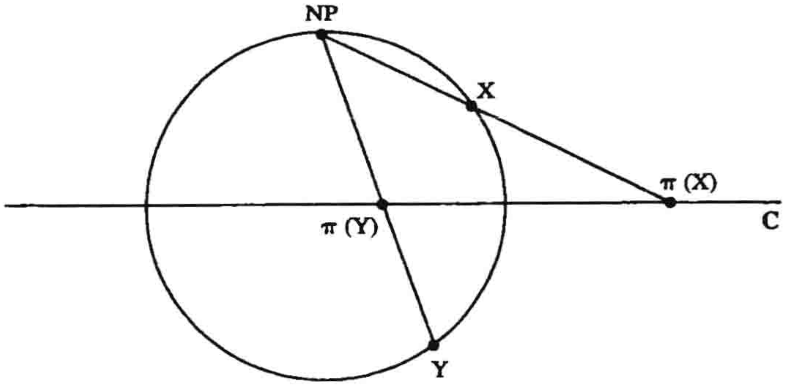


Fig. 1. Stereographic projection $\pi : \mathbf{S}^2 \to \Re$ *is a bijective map from the real 2-sphere to the Riemann sphere* $\Re \cong \mathbf{C} \cup \{\infty\}$. *The 'North Pole' maps to* $\infty \in \Re$.

2. The 'Dodecahedral Set' of 40 rays

2.1. 20 'explicit' rays

In defining the Dodecahedral set of rays, one makes use of the *Majorana map* (Majorana 1932), which is a bijective correspondence between the set of rays in complex n-dimensional Hilbert space $\mathbf{H}^n$ and the set of unordered $(n-1)$-tuples of points on the real sphere $\mathbf{S}^2 \subset \mathbf{R}^3$ (see Penrose 1992). In the present context, we are interested in representing rays in $\mathbf{H}^4$ as unordered triples of points on the sphere. As we shall see, this allows us to make use of some interesting and easily-visualized geometry.

For our purposes, the Majorana map may be defined in the following way. Let d be a ray in complex four-dimensional Hilbert space $\mathbf{H}^4$, and let $\mathbf{d}$ be a (non-zero) representative vector, that is, one that spans d. If we choose a basis for $\mathbf{H}^4$, then $\mathbf{d}$ has components

$$\mathbf{d} = \begin{pmatrix} \Delta_0 \\ \sqrt{3}\Delta_1 \\ \sqrt{3}\Delta_2 \\ \Delta_3 \end{pmatrix},$$

where $\Delta_i \in \mathbf{C}$. (The unusual definitions for Δ_1 and Δ_2 have been made for convenience in what follows.)

Using the Δ_i, form a polynomial

$$\Delta(z) = \Delta_3 z^3 + 3\Delta_2 z^2 + 3\Delta_1 z + \Delta_0.$$

If $\Delta_3 \neq 0$, then this polynomial will have three roots $d_1, d_2, d_3 \in \mathbf{C}$. And in any case, if the leading p coefficients are zero, then the polynomial will still have three roots in the *Riemann sphere* $\Re \cong \mathbf{C} \cup \{\infty\}$, provided we set $d_1 = \ldots = d_p = \infty$. Making use of stereographic projection from the Riemann sphere to the real 2-sphere $\mathbf{S}^2 \subset \mathbf{R}^3$ (see Fig. 1), we have therefore associated with the original ray d an unordered triple of points (D_1, D_2, D_3) on the sphere. Informally then, the (bijective) Majorana map is

$$d \longleftarrow\!\!-\!\!\longrightarrow \Delta(z) \longleftarrow\!\!-\!\!\longrightarrow d_i \in \mathfrak{R} \longleftarrow\!\!-\!\!\longrightarrow D_i \in \mathbf{S}^2,$$

where the points $D_i \in \mathbf{S}^2$ project stereographically to the roots $d_i \in \mathfrak{R}$ of $\Delta(z)$. We denote by $[d]$ the triple (D_1, D_2, D_3) of points representing d.

As an example of the Majorana map, consider the ray d spanned by the vector

$$\mathbf{d} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

which is the $m = +\frac{1}{2}$ eigenstate of spin in the $\hat{z}$ direction. The polynomial $\Delta(z)$ corresponding to d is

$$\Delta(z) = \sqrt{3}z,$$

which has roots ∞, ∞, $0 \in \mathfrak{R}$. These points are the stereographic projections of the points NP, NP, SP, where NP and SP are the 'north and south poles' of the sphere. Thus $[d] = (NP, NP, SP)$. And in general, the $m = +\frac{1}{2}$ eigenstate of spin, in the direction specified by a point X of the sphere with centre at the origin O, will have Majorana representation $(X, X, -X)$, where the minus sign represents reflection in the centre of the sphere.

Now imagine a regular dodecahedron inscribed within the unit sphere with centre at O. Label the vertices of the dodecahedron $A, B, \ldots, M, N, P, Q, \ldots, U$ as in Fig. 2. These 20 points determine 20 triples of the form $(X, X, -X)$, and so they determine 20 $m = +\frac{1}{2}$ spin eigenstates, or rays in $\mathbf{H}^4$. These are the so-called 'explicit' rays in Penrose's non-colourable set.

Soon we shall define the other 20 rays in the set (the 'implicit' rays). But before we can do this, we must first describe some of the ways in which orthogonality relations among rays relate to adjacency relations among corresponding vertices of the dodecahedron.

2.2. The first three orthogonality rules

Consider the ray d from the last section, which is the $m = +\frac{1}{2}$ eigenstate of spin in the $\hat{z}$ direction. This ray has Majorana representation (NP, NP, SP). By looking at any representative vector for d, we see that a ray g will be orthogonal to d if and only if its representative vectors have zeroes for their second components. In the notation of the last section, we must have

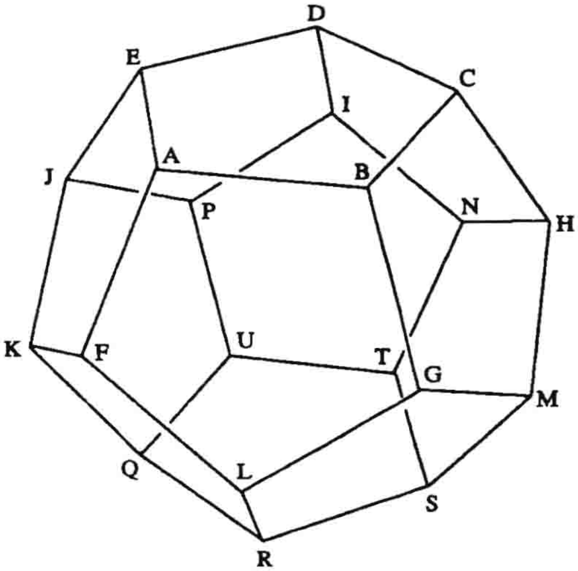


Fig. 2. A regular dodecahedron, imagined to be inscribed in the unit sphere $\mathbf{S}^2 \subset \mathbf{R}^3$ *with centre at the origin* O. *(Output from the computer program* Mathematica.*)*

$$\mathbf{g} = \begin{pmatrix} \Gamma_0 \\ \sqrt{3}\Gamma_1 \\ \sqrt{3}\Gamma_2 \\ \Gamma_3 \end{pmatrix},$$

where $\Gamma_1 = 0$. In terms of the characteristic polynomial $\Gamma(z)$, we have

$$\Gamma(z) = \Gamma_3 z^3 + 3\Gamma_2 z^2 + \Gamma_0.$$

Let us specialize to the case where $\Gamma_3 \neq 0$. If we denote by g_1, g_2, g_3 the roots of $\Gamma(z)$, then it is easy to show that the lack of a z term in this polynomial is equivalent to the condition that

$$g_1 g_2 + g_1 g_3 + g_2 g_3 = 0.$$

This is therefore the condition for $<d|g> = 0$ (in the case $\Gamma_3 \neq 0$).

This condition will become easier to work with if we make use of the fact that if a point $G \in S^2 \setminus \{NP, SP\}$ has projection g, then its reflection $-G$ will have projection $-1/\bar{g}$. Taking this into account, it is easy to see that the condition $g_1 g_2 + g_1 g_3 + g_2 g_3 = 0$ on the triple (G_1, G_2, G_3) is equivalent to the condition $(-1/\bar{g}_1) + (-1/\bar{g}_2) + (-1/\bar{g}_3) = 0$ on the *reflected* triple $(-G_1, -G_2, -G_3)$. This means that the triples (NP, NP, SP) and (G_1, G_2, G_3) represent orthogonal rays if the projections of the points in the reflected triple $(-G_1, -G_2, -G_3)$ add to zero.

Considering once again the special case $(G_1, G_2, G_3) = (G, G, -G)$, we have as the orthogonality condition that the projections of the points $(-G, -G, G)$

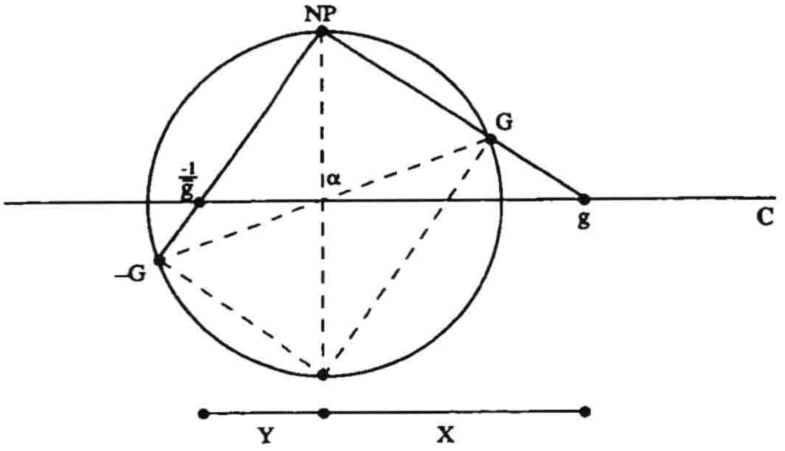


Fig. 3. If g *is the projection of* G, *then* $-1/\bar{g}$ *is the projection of* $-$G. *By using similar triangles, one can show that* $(-1/\bar{g})+(-/\bar{g})+(g) = 0$, *i.e.* X = 2Y, *if* $cos(\alpha) = \frac{1}{3}$.

must add to zero. It is not hard to show that this will happen when the angle subtended by NP and G is $\alpha \equiv cos^{-1}(\frac{1}{3})$ (see Fig. 3). What is more, the condition $<a|b> = 0$ between rays corresponds to a condition between triples of points on the sphere that is invariant under rotations (Penrose 1992). Given what has just been said, this means that any two triples $(X, X, -X)$ and $(Y, Y, -Y)$ represent orthogonal rays when the angle subtended by X and Y is α. Conveniently, this is the angle subtended by adjacent vertices of a cube (see Fig. 4). Still more conveniently, the dodecahedron itself contains *five* cubes, which sit in such a way that any two next-to-adjacent vertices of the dodecahedron are adjacent vertices of some cube (see Fig. 5). Thus we have the following orthogonality rule:

I. Explicit rays x and y, corresponding to next-to-adjacent vertices X and Y of the dodecahedron, are orthogonal.

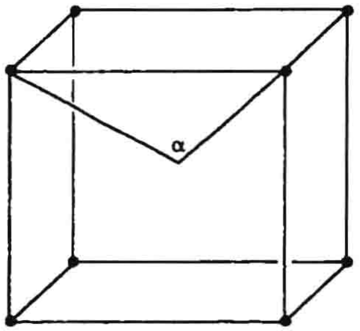


Fig. 4. The angle α *between adjacent vertices of the cube is* $cos^{-1}(\frac{1}{3})$. *This means that the rays* x *and* y *with Majorana representations* (X, X, $-$X) *and* (Y, Y, $-$Y) *are orthogonal when* X *and* Y *are adjacent vertices of the cube.*

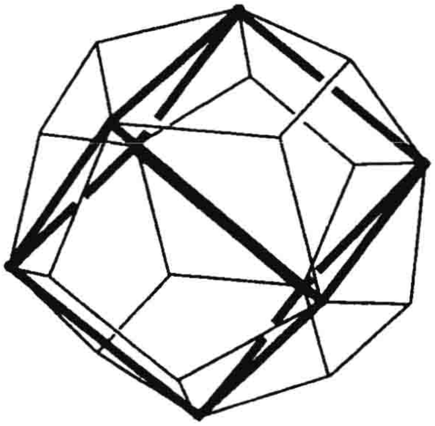


Fig. 5. One of the five cubes contained in the dodecahedron. Any two next-to-adjacent vertices of the dodecahedron are adjacent vertices of some cube, so the rays corresponding to these vertices are orthogonal. (Mathematica).

The second orthogonality rule is a consequence of the fact that we are working in a Hilbert space of even dimension. It is the following:

> II. Rays g and $\tilde{g}$, whose Majorana representations $[g]$ and $[\tilde{g}]$ are reflections of one another in the centre of the sphere, are orthogonal.

To show this, let g and $\tilde{g}$ be rays whose Majorana representations $[g] = (G_1, G_2, G_3)$ and $[\tilde{g}] = (-G_1, -G_2, -G_3)$ are reflections of one another in the centre of the sphere. Remember that if the projections of the points G_i are denoted g_i, then the projections of the points $-G_i$ will be equal to $-1/\overline{g}_i$, where the overbar denotes conjugation.

Now, it is easy to see that if the g_i are the roots of a polynomial

$$\Gamma(z) = \Gamma_3 z^3 + 3\Gamma_2 z^2 + 3\Gamma_1 z + \Gamma_0$$

then the $-1/\overline{g}_i$ will be roots of

$$\hat{\Gamma}(z) = -\overline{\Gamma}_0 z^3 + 3\overline{\Gamma}_1 z^2 - 3\overline{\Gamma}_2 z + \overline{\Gamma}_3.$$

And these are the characteristic polynomials of rays spanned by the vectors

$$\begin{pmatrix} \Gamma_0 \\ \sqrt{3}\Gamma_1 \\ \sqrt{3}\Gamma_2 \\ \Gamma_3 \end{pmatrix}$$

and

$$\begin{pmatrix} \overline{\Gamma}_3 \\ -\sqrt{3}\,\overline{\Gamma}_2 \\ \sqrt{3}\,\overline{\Gamma}_1 \\ -\overline{\Gamma}_0 \end{pmatrix}.$$

which have inner product $\Gamma_0\Gamma_3-3\Gamma_1\Gamma_2+3\Gamma_2\Gamma_1-\Gamma_3\Gamma_0\equiv 0$. Thus orthogonality rule (II) is established.

2.3. 20 'implicit' rays

We shall use rules (I) and (II) often in what follows, but for the moment we consider only rule (I). Notice that the three dodecahedron vertices X_1, X_2, X_3, that are adjacent to a given vertex X, are all next-to-adjacent to one another. Thus rule (I) implies that the three rays x_1, x_2, x_3 corresponding to them are mutually orthogonal, and they determine a unique ray x' that completes the orthogonal tetrad. We refer to x' as the 'implicit' ray corresponding to X. In this way, the 20 vertices of the dodecahedron determine a set **Q** of 40 rays in $\mathbf{H}^4$.

Notice that because of the way the implicit rays have been defined, we have, in addition to (I) and (II), the third orthogonal rule:

III. The explicit ray x and the implicit ray y', corresponding to adjacent vertices X and Y of the dodecahedron, are orthogonal.

3. Contradiction with local deterministic hidden variables

3.1. Rules for labelling rays

In this Section we show how the 40 rays in **Q** can be used to give a proof of Bell non-locality without probabilities (as done originally in Penrose 1992). However, it will be helpful if we first review the basic idea behind proofs of the inconsistency of *non-contextualist* hidden-variables theories. From there, in keeping with the work of Heywood and Redhead, it will be a small step to consider *local* hidden-variable theories.

Let a particular system S be specified. In a hidden-variable theory, the result of any measurement performed on S will depend only upon the (complete) state of S. Given this state, therefore, the result of a particular measurement must be 'decided in advance' of that particular measurement.

If the theory is to be non-contextual as well, then its predictions for the outcomes of measurements must not depend upon the particular contexts in which the measurements are made. For instance, suppose we have a set of distinct physical quantities $Q=\{Q_1, Q_2, \ldots, Q_k\}$. Suppose further that Q has subsets R and T, sharing a quantity Q_i, such that the quantities in R can be measured together in a single context C_R, and the quantities in T can be measured together in a single context C_T, but the quantities in $R\cup T$ cannot all be measured together in the same context. In this circumstance, measuring Q_i involves choosing between the contexts C_R and C_T. To be non-contextualist, the hidden-variable theory must give a prediction for the outcome of the measurement of Q_i that is independent of this choice.

Let us now require that the non-contextualist hidden-variable theory be consistent with standard quantum mechanics. In standard quantum mechanics the system S may be represented (however incompletely) by a vector in some Hilbert space, and measurements on S may be represented by operators on this space. Thus in a 'consistent' hidden-variable theory,[4] a given operator P has assigned to it a fixed number $f(P)$, equal to what the outcome of the corresponding measurement on S would be, were it performed. In addition, the theory is said to be *non-contextualist* if the 'label' $f(P)$ is independent of the context in which the measurement corresponding to P is made.

The possible outcomes of the measurement corresponding to P are represented in the quantum mechanical formalism by the eigenvalues of P. For convenience, therefore, we may consider the non-contextual hidden-variable labelling f to be a map $f: P \mapsto \lambda \in \{\text{eigenvalues of } P\}$. With this convention, the non-contextualist assumption is that the labelling f is a well-defined function of P only, even when P is degenerate.

We can obtain a constraint on the hidden-variable labelling f by considering measurements corresponding to the one-dimensional projection operators on n-dimensional Hilbert space $\mathbf{H}^n$. These operators are of the form $P_\psi = \psi\psi^\dagger$, where ψ is a normalized state vector. The eigenvalues of such an operator are 1 and 0, and they correspond to "yes" and "no" answers to the question, "Is the system in the state ψ?"

Now suppose that there is a set $\{\psi_1, \ldots, \psi_n\}$ of mutually orthonormal states in $\mathbf{H}^n$, with associated operators $P_i = \psi_i\psi_i^\dagger$ projecting to them. These operators commute, so the measurements corresponding to them can be performed in a single context. In this set of measurements, the outcomes correspond to "yes" and "no" answers; and among these n answers, there must be one and only one "yes". The fact that there must be a "yes" answer places the following constraint on the non-contextual hidden-variable labelling f:

> *for any n mutually orthonormal states in $\mathbf{H}^n$, it cannot happen that all n of the associated projection operators are labelled 0 by a consistent, non-contextualist, deterministic hidden-variable theory.*

And the fact that there must be *only one* "yes" answer places a second constraint on the hidden-variable labelling:

> *if ψ_1 and ψ_2 are orthonormal states, then the operators P_1 and P_2 projecting to them cannot both be labelled 1 by a consistent, non-contextualist, deterministic hidden-variable theory.*

---

[4] A 'consistent' hidden-variable theory is one that is consistent with the predictions of standard quantum mechanics.

Otherwise it would be possible to choose $n-2$ states that combine with ψ_1 and ψ_2 to form a mutually orthonormal set, and then the labelling would contradict the quantum predictions for the complete set of measurements projecting to these states.

These constraints are actually specific instances of a more general restriction that one can place on the hidden-variable labelling. This restriction is the so-called "Sum Rule" (see Redhead 1987) and it may be stated as follows. Let A and B be a pair of commuting operators representing measurement processes. In standard quantum mechanics, the operator $A+B$ will also represent a physical observable, one which can be measured by measuring A and B and adding the outcomes. That is to say, in standard quantum mechanics, the outcome of the simultaneous measurement of $A+B$ is the same as the sum of the outcomes of consecutive measurements of A and of B performed in the same context. For this reason, it is a necessary feature of a consistent hidden-variable theory that the label assigned to $A+B$ be the sum of the labels assigned to A and to B. Succinctly then, the Sum Rule states that $f(A+B)=f(A)+f(B)$ for the hidden-variable labelling f.

Now suppose that $\{\psi_1, \ldots, \psi_n\}$ is an orthonormal set of states in $\mathbf{H}^n$. It is easy to see that the commuting operators $P_i = \psi_i \psi_i^\dagger$ projecting to these states will satisfy the equation

$$P_1+P_2+\cdots+P_n = I$$

where I is the identity operator. Making use of the Sum Rule, therefore, we have an equation in the operators' labels:

$$f(P_1)+f(P_2)+\cdots+f(P_n)=f(I).$$

And since $f(P_i) \in \{0,1\}$ and $f(I) = 1$, this equation has the implications given above in italics.

The two constraints listed above provide a conceptually simple way of proving the Bell–Kochen–Specker Theorem: if there is a set of projection operators, no possible labelling of which satisfies the two italicized requirements, then no non-contextualist deterministic hidden-variable theory can be consistent with all of the predictions of standard quantum mechanics. As mentioned previously, Kochen and Specker gave such a proof using 117 operators in $\mathbf{H}^3$.

The above considerations have been phrased in terms of operator labellings, but of course one may just as well speak of labelling the rays to which these operators project. Phrased in these terms, and specialized for the present case of $n = 4$, the conditions that a non-contextualist deterministic hidden-variable labelling must satisfy are

(3.1) No two orthogonal rays are both labelled 1.

(3.2) No four mutually orthogonal rays are all labelled 0.

Up to now we have spoken only of *non-contextuality* in hidden-variable theories. But as we will show next, the rules (3.1) and (3.2) are also the predictions of a consistent *local* hidden-variable theory, for either of two particles in an EPR arrangement. To see this, consider the following argument.[5]

Imagine that we have an EPR pair of spin-3/2 particles emitted to the right and to the left in the singlet state. Standard quantum mechanics predicts that a measurement on the *left* particle, projecting onto the ray ψ with Majorana representation $[\psi]$, will have the same result as a measurement on the ***right*** particle that projects onto the ray whose Majorana representation is the *reflection* of $[\psi]$ in the centre of the sphere. (For convenience, let us write $-[\psi]$ to denote the reflection of $[\psi]$ in the centre of the sphere.) Following Penrose (1992), we can use this anticorrelation, together with the locality assumption, to establish that there must in fact be a non-contextual labelling of the set of all rays (or equivalently, of the operators projecting to those rays).

To do this, let $f_L(\psi) \in \{0,1\}$ denote the label assigned to the ray ψ by the hidden-variable theory (giving the outcome of a measurement on the left particle projecting onto ψ). Then anticorrelation implies that there is a label $f_R(\tilde{\psi})$, on the ray $\tilde{\psi}$ with Majorana representation $-[\psi]$, which must be the same (the subscript R refers to the right particle); that is, $f_R(\tilde{\psi}) = f_L(\psi)$. Under the assumption of locality, making a measurement on the left particle could not have caused this label to come into existence, so it must have existed previously. In this way, we see that each ray ϕ has a labelling $f_R(\phi) \in \{0,1\}$ that gives the outcome of a measurement on the right particle projecting to ϕ.

To see that this labelling is non-contextual, we need only note that the anticorrelation rule is phrased without reference to the measurement context. So even if we presume for the moment that local context matters, and we say that the label $f_L(\phi)$ is defined with respect to a particular context, the anticorrelation rule says that $f_R(\tilde{\psi}) = f_L(\psi)$ regardless of the context in which the right-hand label is defined. Thus the right particle has a non-contextual labelling f_R—and by symmetry, the left particle does as well. This means that we may appeal once again to the Sum Rule to see that the rules (3.1) and (3.2) apply to each of the two particles.

3.2. No labelling exists satisfying (3.1) and (3.2)

To establish Bell's Theorem, therefore, it need only be shown that there is no labelling of the rays in **Q** satisfying (3.1) and (3.2). To do this, we show that these ray-labelling rules give rise a couple of similar rules for colouring vertices of the dodecahedron. This will allow us to appeal to a straightforward geometric argument to show that (3.1) and (3.2) cannot be satisfied for **Q**.

---

[5] Some portions of this following argument are from Penrose (1992); we are grateful to Dr Harvey Brown for helping to construct the others.

A ray labelling $f\colon \mathbf{Q} \to \{0,1\}$ gives rise to a natural colouring f^* of the vertices of the dodecahedron, defined by

$$f^*(X) = \textit{black}, \text{ if } f(x) = 0$$
$$f^*(X) = \textit{white}, \text{ if } f(x) = 1$$

where x is the ray for which $[x] = (X, X, -X)$.

If a labelling $f\colon \mathbf{Q} \to \{0,1\}$ satisfies (3.1) and (3.2), then it is easy to see that f^* must satisfy the following:

D.1. No two next-to-adjacent vertices are both coloured *white*.

D.2. The six vertices adjacent to one or the other of a pair of antipodal vertices are not all coloured *black*.

It is obvious that rule (D.1) follows from (3.1), since next-to-adjacent vertices represent orthogonal rays. To see that (D.2) follows from (3.1) and (3.2), let X be a vertex of the dodecahedron, and imagine that the six vertices adjacent to X and $-X$ have all been coloured black by some colouring f^*. In particular, then, the three vertices X_1, X_2, and X_3 adjacent to X are coloured black. Therefore in the corresponding colouring $f\colon \mathbf{Q} \to \{0,1\}$, we must have $f(x_1) = f(x_2) = f(x_3) = 0$, where x_i is the ray for which $[x_i] = (X_i, X_i, -X_i)$. But these three rays are mutually orthogonal (by (I)), so by (3.2) we must have $f(x') = 1$, where x' is the implicit ray that completes the tetrad. And by a similar argument, if $-X_1$, $-X_2$, and $-X_3$ are all coloured black, then we must have $f(\tilde{x}') = 1$ as well, where $\tilde{x}'$ is the implicit ray corresponding to $-X$. But now x' and $\tilde{x}'$ are orthogonal (by (II)), so by (3.1) they *cannot* both be labelled 1. This contradiction shows that (D.2) follows from (3.1) and (3.2).

To complete the proof, we note that there is in fact no colouring of the vertices of the dodecahedron, with the colours black and white, that satisfies (D.1) and (D.2). It is relatively straightforward to show this, so in the interest of space the reader is referred to Penrose (1992), where the argument was given originally.

From the impossibility of such a colouring, we may conclude that the set **Q** provides a geometrically intriguing proof, free of probabilities, of Bell's Non-Locality Theorem.

4. The full orthogonality relations

4.1. The last three orthogonality rules

For the purposes of extracting a non-colourable subset from **Q**, it is desirable to have a complete account of the orthogonality relations holding among these rays. Together with (I)–(III), the following three rules give us such an account:

IV. The explicit and implicit rays x and x', corresponding to a vertex X of the dodecahedron, are orthogonal.

V. The explicit and implicit rays x and $\tilde{x}'$, corresponding to antipodal vertices X and $-X$, are orthogonal.

VI. The implicit rays x' and y', corresponding to next-to-next-to-adjacent vertices X and Y of the dodecahedron, are orthogonal.

Because of rotational invariance, one can prove (IV) by showing that d and d' are orthogonal, where once again d is the $m = +\frac{1}{2}$ eigenstate of spin in the $\hat{z}$ direction. As in the discussion of orthogonality rule (I), this will be true if the projections of the points $-D_i'$ of $-[d']$ add to zero. This is in fact the case, as these points are equally spaced about a circle that is situated normal to the $\hat{z}$ axis (see Zimba 1993). Rule (V) also follows from this circumstance, since the reflected points D_i' of $-[\tilde{d}']$ are also equally spaced about a circle normal to the $\hat{z}$ axis.

As for (VI), the proof given in (Zimba 1993) is a brute force calculation. Here we give a much simpler argument, showing that (VI) is in fact implicit in (I)–(V). We first note (Fig. 2) that, by (I), each of a and b is orthogonal to both of l and d. Moreover, by (III) and (V), each of r' and i' is also orthogonal to both l and d. It follows that r' and i' must each belong to the span of a and b, and are therefore orthogonal to anything that is orthogonal to both of a and b. But by (III) and (IV), each of a and b is orthogonal to a', whence r' and i' must each be orthogonal to a'. These are typical instances of (VI), and the rest follow by rotational symmetry of the dodecahedron.

The rules (I)–(VI) characterize every type of orthogonality among the rays in **Q**. Using these rules, one finds that the set contains 40 orthogonal tetrads: 20 of these consist of three explicit rays and one implicit one, 10 consist of two explicit rays and two implicit ones, and 10 consist of four implicit rays. Each explicit ray is a member of four orthogonal tetrads, and each implicit ray is a member of four orthogonal tetrads as well.

The set **Q** contains all-implicit tetrads because the dodecahedron has groups of four mutually next-to-next-to-adjacent vertices. These correspond to the 10 tetrahedra contained in the dodecahedron (Fig. 6).

The following is a list of the 40 orthogonal tetrads in **Q**. Every orthogonality is represented in the list; that is, the set has the property that no two rays are orthogonal unless they are together in some orthogonal tetrad.

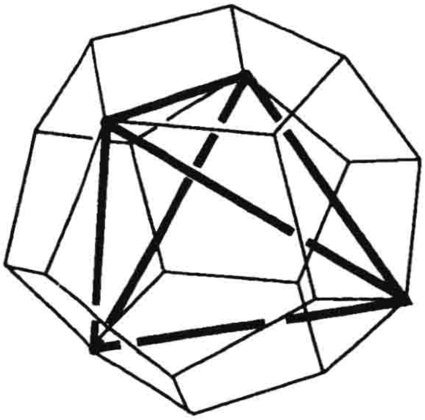


Fig. 6. One of the 10 tetrahedra contained in the dodecahedron, whose corresponding implicit states form an orthogonal tetrad. (Mathematica).

feba'
acgb'
dbhc'
iecd'
jdae'
kalf'
lbmg'
cnmh'
pdni'
pkej'
qfjk'
rfgl'
gshm'
tihn'
juip'
kruq'
lqsr'
rtms'
unst'
pqtu'
aa'tt'
bb'uu'
cc'qq'
dd'rr'
ee'ss'
ff'nn'
gg'pp'
hh'kk'
ii'll'
jj'mm'
a'i'q'm'
a'p'r'h'
b'k'i's'
b'j'n'r'
c'f'p's'
c'l'j't'
d'k'g't'
d'f'm'u'
e'l'h'u'
e'g'n'q'

A further property of the set is that any two tetrads in this list share at most one ray. This allows us to observe, for instance, that Peres's set of 24 rays in $\mathbf{H}^4$ is not a subset of **Q**: for Peres's set contains tetrads with representative vectors

$$(1,0,0,0) \quad (0,1,0,0) \quad (0,0,1,0) \quad (0,0,0,1)$$

and

$$(1,0,0,0) \quad (0,1,0,0) \quad (0,0,1,-1) \quad (0,0,1,1)$$

(see Peres 1991), and these share *a pair* of rays.

4.2. *Observations on the structure of the set*

The fact that there are exactly as many tetrads in **Q** as there are rays leads one to wonder whether the tetrads and rays are in some sense dual to one another. This suspicion is strengthened by the fact that each ray is orthogonal to 12 others, while each tetrad shares a ray with 12 others. Thus if one forms the *Kochen–Specker diagram* of **Q**, which is a graph Q in which vertices represent rays and edges represent orthogonalities, one might suspect that it is isomorphic to a graph Q^* in which vertices represent *tetrads* and edges represent *shared rays*. Certainly both graphs have 40 vertices with 12 edges meeting at each vertex; moreover, both graphs are made up of 40 4-*cliques*, any two of which share at most a single vertex (a 4-clique is a graph with 4 vertices, each pair of which is joined by an edge). However, these similarities aside, the two graphs are actually different. It is not very hard to show that since **Q** is non-colourable, it must be the case that in any subset of 10 rays in **Q**, there is an orthogonal pair. This means that in any subgraph of Q containing 10 vertices, there must be an edge. In contrast, there are *no* edges in the subgraph of Q^* containing the 10 vertices corresponding to the 'antipodal' tetrads in the third column of the table. Thus Q and Q^* are not isomorphic.

Some insights into the geometrical structure of **Q** can be obtained by examining the configuration of 40 points in the complex projective three-space $\mathbf{CP}^3$ that these rays determine. Rays such as a, b, r', i' that, as we have seen, belong to a linear two-space are represented in $\mathbf{CP}^3$ as four collinear points. Likewise, the points describing l, d, a', b' are collinear. The same holds for a, i, q, m, which are all orthogonal to each of c, l, j, t (two tetrahedra inscribed in a cube contained in the dodecahedron). Moreover, a', c', g', u represent collinear points (all of them being orthogonal to each of q', p', t', b). That these are the only four types of collinearity, up to rotational symmetry of the dodecahedron, follows from the fact that if two points are orthogonal, then there is no other point in the configuration with which they are collinear. (This in turn follows from the fact that there are only two points that are each orthogonal to both of a given pair x,y of orthogonal points—these being the two that complete the single tetrad containing x and y.)

The set of points in $\mathbf{CP}^3$ that represent the *explicit* rays forms a configuration whose collinearities are depicted in Fig. 7. The configuration suggests a useful notation for the vertices of a dodecahedron, and this is given in Fig. 8. In this notation, connected vertices of the dodecahedron are given by pairs $(\alpha\beta)$, $(\gamma\delta)$, where α, β, γ, δ are distinct members of the set $\{1, 2, 3, 4, 5\}$ such that if ε is the remaining member of this set, then $\alpha\beta\gamma\delta\varepsilon$ is an *even* permutation of 12345. Note that $(\alpha\beta)$ is *antipodal* to $(\beta\alpha)$, and that $(\alpha\beta)$ is *orthogonal* to $(\lambda\mu)$ if and only if $\alpha = \mu$ or $\beta = \lambda$ (or both). For convenience, we give a translation table for the two vertex notations:

a:(12)	f:(53)	k:(41)	q:(32)
b:(45)	g:(31)	l:(24)	r:(15)
c:(23)	h:(14)	m:(52)	s:(43)
d:(51)	i:(42)	n:(35)	t:(21)
e:(34)	j:(25)	p:(13)	u:(54)

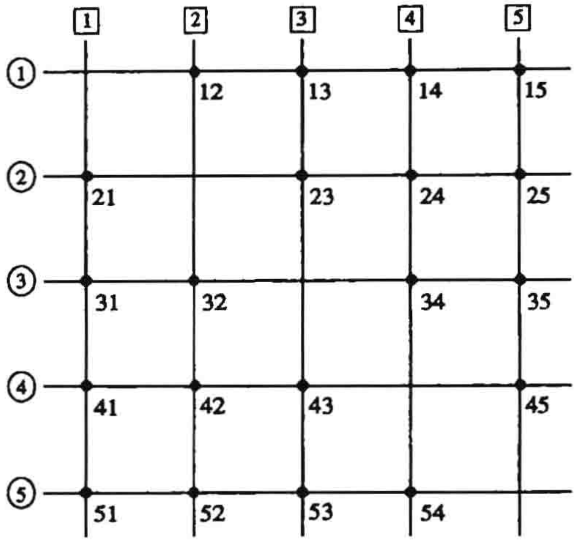


Fig. 7. Diagram showing collinearities in the set of 20 explicit points in $\mathbf{CP}^3$. *Lines whose intersections are not marked with spots are skew. The four points on a given line correspond to a tetrahedron in the dodecahedron.*

If we try to represent the *implicit* rays also on Fig. 7, then the picture becomes messy. However, some good insights as to the total configuration of 40 points are obtained if we consider the system of points orthogonal to a given point of $\mathbf{CP}^3$, say the point (51). This is an interesting configuration, which is not representable on a *real* plane, consisting of 12 points lying by fours on nine lines (see Fig. 9). It is the dual of a configuration known to geometers as the set of nine inflexions of a (complex) general cubic curve (see for example Walker

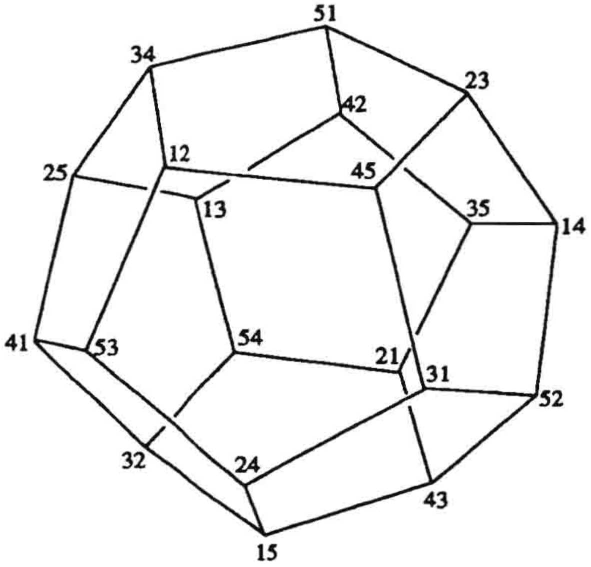


Fig. 8. The vertex notation suggested by the collinearity diagram.

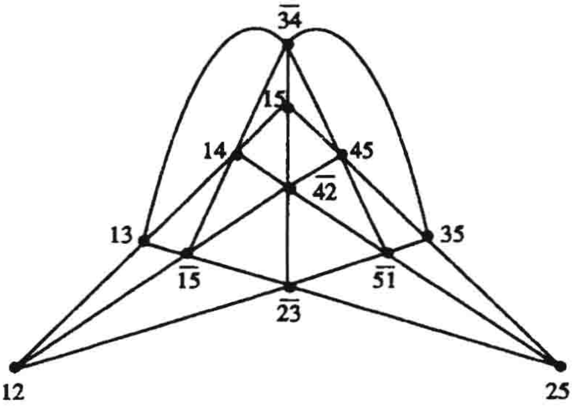


Fig. 9. The 12 points orthogonal to a given point have collinearities as shown. (In this case the 12 points are orthogonal to (51).) Overbars are used to denote implicit points.

(1950), Theorem 6.6, pp. 72,73). Note that any two distinct points in the planar configuration are either (a) orthogonal, or (b) joined by a line containing two other points of the configuration (but not both). Indeed, this holds for any pair of distinct points in the 40-point configuration as a whole in $\mathbf{CP}^3$.

Perhaps the easiest way to see that $\mathbf{Q}$ is essentially complex is to look at the cross-ratio of the four collinear explicit points corresponding to a tetrahedron in the dodecahedron; from this it will follow that the 12-point configuration of Fig. 9 is not representable on a real plane. For instance, consider the tetrahedron (15)(14)(13)(12), whose points lie on one of the lines of Fig. 9. Rotating the dodecahedron by 120 degrees in the left-handed sense about the (12)—(21) axis permutes these points by (15)(14)(13)(12) $\mapsto$(13)(15)(14)(12). This rotation gives rise to a projective transformation on $\mathbf{CP}^3$ effecting the same permutation. Now, if the cross-ratio of four points taken in the order (15)(14)(13)(12) is λ, then the cross-ratio taken in the order (13)(15)(14)(12) is $1/(1-\lambda)$. But a projective transformation preserves cross-ratios, so these two numbers must be the same, that is, we must have $\lambda = 1/(1-\lambda)$, or $\lambda^2-\lambda+1 = 0$. Thus λ is $-\omega$ or $-\omega^2$, where $\omega = e^{2\pi i/3}$ is the cube root of unity. A set of four points with this cross-ratio is said to be *equianharmonic*, and it is clear that there is no projective transformation of an equianharmonic set that gives real coordinates to all the points.

The configuration of 40 points in $\mathbf{CP}^3$ is highly symmetrical. The points lie by fours on 90 lines, with nine lines through each point. The orthogonal complements of the 40 points are 40 planes, each of which contains a configuration of 12 points and nine lines as in Fig. 9. There are 12 such planes through each point, and the configuration of lines and planes through each point has incidence relations given by the dual of Fig. 9, as illustrated in Fig. 10. (This is the aforementioned configuration of nine inflexions of a general cubic curve.) The *points* of Fig. 10 can be thought of as representing lines

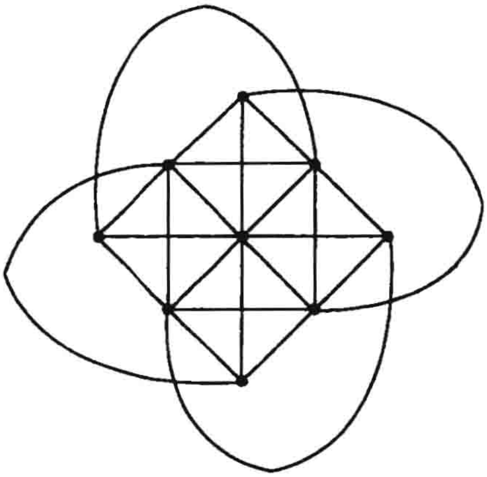


Fig. 10. The projective configuration of the nine inflexions of a general complex cubic curve. This configuration represents the incidence relations of lines and planes through each of the 40 points.

through a given one of the 40 points, and a *line* of Fig. 10 by a plane through it. Think of placing your eye at that point; then the remaining 27 points of the configuration that are not orthogonal to it line up in threes to form the points of Fig. 10, whereas the planes through that point provide the collinearities given in Fig. 10.

Any two of the 40 points are either orthogonal, or else their inner product is $1/\sqrt{3}$ (up to a phase factor). Remembering that there are no orthogonalities amongst the four points on a given line, one sees that the six possible inner products must all have this value. If we represent each line of four points as a Riemann sphere, where orthogonal points are represented as antipodal, then the four points form the vertices of a regular tetrahedron, whose proper rotations can be realized as unitary transformations and improper ones by anti-unitary transformations.

The entire configuration of 40 points in $\mathbf{CP}^3$ can be understood roughly in the following terms. Select any four mutually orthogonal points, representing a tetrad of rays in $\mathbf{H}^4$. Each of the four planes through three of these points contains a configuration identical with that of Fig. 9, giving nine more points (henceforth called 'internal' points) on each plane, and providing, together with the original four points, the entire set of 40. There are nine lines on each of these four planes, giving 36 of the 90 lines of the configuration. The remaining 54 are obtained by joining each of the nine internal points on one of the four planes to each of six internal points of another of these planes. (These are the six points on that plane that are not orthogonal to the chosen point.) Each of these 54 lines cuts the four planes in an equianharmonic set.

Finally, there is a transitive group of 25,920 unitary projective tranformations of $\mathbf{CP}^3$ sending the configuration of 40 points into itself. This is just half

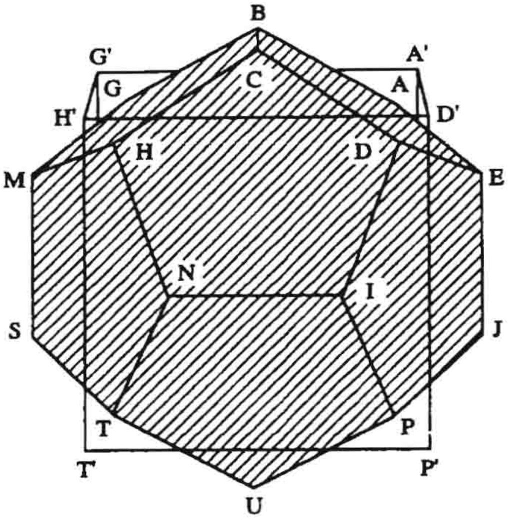


Fig. 11. A visual representation of the states of $\mathbf{Y}$. *The 20 explicit states in* $\mathbf{Y}$ *correspond to the vertices of the dodecahedron; the eight implicit states correspond to the vertices of the cube.* (Mathematica).

of the symmetries of the Kochen–Specker graph of the points, the remaining symmetries being antiunitary transformations (interchanging the cross-ratios $-\omega$ and $-\omega^2$).

5. A non-colourable subset with 28 rays

In this Section we show that there is a non-colourable proper subset $\mathbf{Y} \subset \mathbf{Q}$, whose geometry is based on the dodecahedron and the cube. Specifically, $\mathbf{Y}$ contains the 20 explicit rays corresponding to the vertices of the dodecahedron, as well as the eight implicit rays corresponding to any one of the five possible cubes (Fig. 11).

To see that $\mathbf{Y}$ is non-colourable, suppose contrariwise that there *is* a function $f:\mathbf{Y} \rightarrow \{0,1\}$ labelling the rays in $\mathbf{Y}$ according to (3.1) and (3.2). All 20 of the explicit rays are in $\mathbf{Y}$, so we can define f^* to be the colouring of the vertices of the dodecahedron that corresponds to f.

In keeping with the dodecahedron colouring argument of Section 3, f^* must violate (D.1) or (D.2). However, notice that if f^* violated (D.1), then f would violate (3.1); hence f^* must violate (D.2). That is, there must be some vertex X such that the six vertices adjacent to X and $-X$ are all labelled 0.

Let x be the ray corresponding to X. If x' is in $\mathbf{Y}$, then so is $\tilde{x}'$, and we have the desired contradictions, since in that case f cannot satisfy (3.1) and (3.2) (as we saw previously when we derived (D.2)). On the other hand, suppose x is one of the rays for which x' is *not* in $\mathbf{Y}$ (these are b, c, e, f, i, j, l, m, n, q, s, u). Because of the symmetry of **Y**, all of these rays have the same status, so we can actually assume without loss of generality that x is any one of them.

So suppose $x = i$. This means that the rays n, p, d, f, g, and r are all labelled zero. Considering the tetrad $gg'pp'$, this means that one of g' or p' is labelled 1; by symmetry, we can suppose without loss of generality that it is g'. Then m, b, and d' must be labelled 0 ($lbmg'$ and $d'g'k't'$). Hence r' must be labelled 1 ($dd'rr'$), which implies that h' and a' are labelled 0 ($a'p'r'h'$). But now e and c must be labelled 1 ($feba'$ and $cnmh'$), which contradicts (3.1) ($iecd'$).

Having thus derived a contradiction both when x' is in **Y** and when x' is not in **Y**, we conclude that there is no labelling of the rays in **Y** satisfying (3.1) and (3.2). Therefore $\mathbf{Y} \subset \mathbf{Q}$ is non-colourable, so that it provides a somewhat tighter proof of Bell's Theorem than does **Q** itself.

It is easy to show that **Y** is critical, in the sense that all of its proper subsets are colourable. Clearly every subset of a colourable set is colourable, so it suffices to show that deleting a single ray from **Y** yields a colourable set. And because of the symmetry of **Y** there are only three cases to consider: deleting an implicit ray, deleting an explicit ray x such that x' is in **Y**, and deleting an explicit ray x such that x' is not in **Y**.

In the first case we can assume without loss of generality that the ray deleted is k'. Then one can check that (3.1) and (3.2) are satisfied by the labelling in which r', g', f, c, t, and u all receive the value 1. (Any ray not mentioned is understood to have been assigned the value 0.)

In the second case we can assume without loss of generality that the ray deleted is d. Then (3.1) and (3.2) are satisfied by the labelling that assigns the value 1 to the rays g', a', i, n, q, and k.

Finally, in the third case we can assume without loss of generality that the ray deleted is e. Then (3.1) and (3.2) are satisfied by the labelling that assigns the value 1 to the rays p', t', m, q, d, and k. (Notice that each ray which used to be in an orthogonal tetrad with e has received the value 0. Intuitively, removing e has made these zeros available for "insulating" the necessary 1 values).

6. Finite non-colourable sets in higher dimensions

In this Section we shall show that there are finite non-colourable sets in all finite dimensions $n \geq 3$, except possibly for the case $n = 5$. Since we are concerned with various dimensions, it is convenient to re-state the hidden-variable labelling rules in their general form. The rules are

(6.1) No two orthogonal rays are both labelled 1.

(6.2) In any group of n mutually orthogonal rays, not all of the rays are labelled 0.

Until now we have used the term 'non-colourable' to refer to a set with no such labelling. But at this point we must require that the terminology reflect the dimension at hand, so from now on we will use the term 'non-n-colourable,' which refers to the dimension of the space.

Let

$$U = \{(u_1^1, \ldots, u_n^1), \ldots, (u_1^p, \ldots, u_n^p)\}$$

and

$$V = \{(v_1^1, \ldots, v_m^1), \ldots, (v_1^q, \ldots, v_m^q)\}$$

be two sets of vectors representing non-colourable sets of rays **U** and **V**. (**U** is non-n-colourable, and **V** is non-m-colourable.) Form the set

$$U{*}V = \{\tilde{u}^1, \ldots, \tilde{u}^p, \tilde{v}^1, \ldots, \tilde{v}^q\},$$

where

$$\tilde{u}^i = (u_1^i, \ldots, u_n^i, 0, \ldots, 0)$$

(with m 0s), and

$$\tilde{v}^j = (0, \ldots, 0, v_1^j, \ldots, v_m^j)$$

(with n 0s) (see Fig. 12). Then we claim that the corresponding set of rays $\mathbf{U{*}V}$ is non-$(n+m)$-colourable.

The rules (6.1) and (6.2) for n-colouring sets of rays are obviously equivalent to the following two rules for n-colouring sets of representative vectors:

(6.1′) If two vectors are orthogonal, then the product of their labels is 0.

(6.2′) In any subset of n mutually orthogonal vectors, the sum of the labels is greater than 0. (For simplicity, call n mutually orthogonal vectors an n-basis.)

Therefore we can show that $\mathbf{U{*}V}$ is non-$(n+m)$-colourable by (6.1) and (6.2) by establishing that $U{*}V$ is non-$(n+m)$-colourable by (6.1′) and (6.2′). To do this, we show that any $(n+m)$-colouring $f : U{*}V \to \{0,1\}$ satisfying (6.1′) and 6.2′) gives rise naturally to n- and m-colourings $f_1 : U \to \{0,1\}$ and $f_2 : V \to \{0,1\}$, one of which must satisfy (6.1′) and (6.2′). Thus if there are no admissible colourings for U or for V, there cannot be an admissible colouring for $U * V$.

So suppose $f : U * V \to \{0,1\}$ satisfies (6.1′) and (6.2′). Define $f_1 : U \to \{0,1\}$ and $f_2 : V \to \{0,1\}$ by

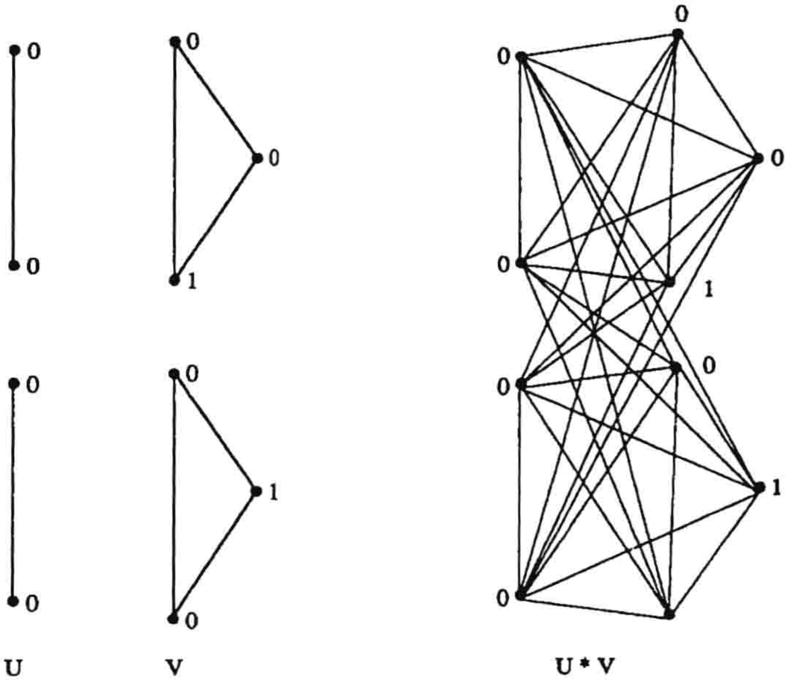


Fig. 12. An example of combining sets via the $*$ operation.

$$U = \{(0,1),\quad (1,0),\quad (1,1),\quad (1,-1)\}$$

$$V = \{(1,0,0),\quad (0,1,0),\quad (0,0,1),\quad (1,1,1),\quad (1,2,-3),\quad (-5,4,1)\}$$

$$\begin{aligned} U_*V = \quad & (0,1,0,0,0),\quad (1,0,0,0,0),\quad (1,1,0,0,0),\quad (1,-1,0,0,0),\\ & (0,0,1,0,0),\quad (0,0,0,1,0),\quad (0,0,0,0,1),\quad (0,0,1,1,1),\\ & (0,0,1,2,-3),\quad (0,0,-5,4,1)\} \end{aligned}$$

U_*V has an admissible colouring, shown on the Kochen–Specker diagram. This induces the colourings on U and on V. The colouring on V is admissible.

$$f_1\ (u^i) = f(\tilde{u}^{ji})$$
$$f_2\ (v^j) = f(\tilde{v}^j).$$

To see that f_1 satisfies (6.1′), let u^i and u^j be two orthogonal vectors in U. Then $\tilde{u}^i$ and $\tilde{u}^j$ are orthogonal in $U * V$, so since f satisfies (6.1′), we must have $f(\tilde{u}^i f(\tilde{u}^j) = 0$, which is to say $f_1(u^i)f_1(u^j) = 0$. Thus f_1 satisfies (6.1′). Similarly one sees that f_2 satisfies (6.1′). Further, if $f_1(u_1^i)+\ldots+f_1(u_n^i) > 0$ for each n-basis $\{u^{i_j}\}$ in U, then f_1 satisfies (6.2′) as well, and we are done.

On the other hand, suppose there is a particular n-basis $\{u^{i_1}, \ldots, u^{i_n}\}$ in U such that $f_1(u^{i_1})+\ldots+f_1(u^{i_n}) = 0$. We note that for any m-basis $\{v^{j_k}\}$ in V, the vectors $\{\tilde{u}^{i_1}, \ldots, \tilde{u}^{i_n}, \tilde{v}^{j_1}, \ldots, \tilde{v}^{j_m}\}$ form an $n+m$-basis in $U*V$, so that

$$f(\tilde{u}^{i_1})+\ldots+f(\tilde{u}^{i_n})+f(\tilde{v}^{j_1})+\ldots+f(\tilde{v}^{j_m}) > 0.$$

This means that

$$f_1(u^{i_1})+\ldots+f_1(u^{i_n})+f_2(v^{j_1})+\ldots+f_2(v^{j_m}) > 0,$$

and since $f_1(u^{i_1})+\ldots+f_1(u^{i_n}) = 0$, we must have $f_2(v^{j_1})+\ldots+f_2(v^{j_m}) > 0$. But since this holds for every m-basis $\{v^{j_k}\}$, we have that f_2 satisfies (6.2′), and the assertion is proved.

This result means that one can combine any two existing non-colourable sets, one containing p rays in $\mathbf{H}^n$, the other containing q rays in $\mathbf{H}^m$, and obtain a new non-colourable set containing $p+q$ rays in $\mathbf{H}^{n+m}$. Since there are existing sets in $\mathbf{H}^3$ and in $\mathbf{H}^4$, we can therefore create new sets in any dimension k as long as k is the sum of some number of 3s and some number of 4s. But it is easy to show that any number $k > 5$ satisfies this condition, by considering the cases $k = 3s$, $k = 3s+1$, and $k = 3s+2$. (For example, when $k = 3s+1$, we have $k = 3(s-1)+3+1 = 3(s-1)+4(1)$.) Thus these are finite non-colourable sets of rays in every Hilbert space with finite dimension greater than 2, except possibly for $\mathbf{H}^5$.

Of course, one suspects that there is such a set in $\mathbf{H}^5$ as well!

Acknowledgements—We would like to thank Dr Harvey Brown and Professor Michael Redhead for suggesting improvements in our presentation of the material in Section 3.1. Dr Brown deserves additional thanks as well, for engaging us in frequent and fruitful discussions of many aspects of this topic. Funding for this research was provided by The Rhodes Trust.

References

Bell, J. S. (1966) 'On the Problem of Hidden Variables in Quantum Theory', *Reviews of Modern Physics* **38**, pp. 447–452.

Brown, H. R. (1991) 'Bell's Other Theorem and its Connection with Nonlocality, Part I, forthcoming in *Bell's Theorem and the Foundations of Physics*, eds. A. Van der Merwe and F. Selleri (Singapore: World Scientific, 1993).

Brown, H. R. and Svetlichny, G. (1990) 'Nonlocality and Gleason's Lemma, Part I: Deterministic theories', *Foundations of Physics* **20** (11), pp. 1379–1387.

Clifton, R. (1991) 'Getting Contextual and Nonlocal Elements-of-reality the Easy Way', forthcoming in *American Journal of Physics*.

Gleason, A. (1957) 'Measures on the Closed Subspaces of a Hilbert Space', *Journal of Mathematics and Mechanics* **6**, pp. 885-893.

Greenberger, D. M., Horne, M. and Zeilinger, A. (1989) 'Going beyond Bell's Theorem', in M. Kafatos, ed., *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (Dordrecht: Kluwer Academic Press).

Greenberger, D. M., Horne, M., Shimony, A. and Zeilinger, A. (1990) 'Bell's Theorem without Inequalities', *American Journal of Physics* **58** (12).

Heywood, P. and Redhead, M. L. G. (1983) 'Nonlocality and the Kochen–Specker Paradox', *Foundations of Physics* **13** (5), pp. 481–499.

Kochen, S. and Specker, E. P. (1967) 'The Problem of Hidden Variables in Quantum Mechanics', *Journal of Mathematics and Mechanics* **17**, pp. 59–88.

Majorana, E. (1932) 'Atomi Orientati in Campo Magnetico Variabile', *Nuovo Cimento* **9**, pp. 43 50.

Penrose, R. (1992) 'On Bell Non-locality Without Probabilities: some Curious Geometry', preprint, Mathematical Institute, Oxford (to appear in *Quantum Reflections*, eds. J. Ellis and D. Amati, Cambridge University Press, Cambridge).
Peres, A. (1990) 'Incompatible Results of Quantum Measurements', *Physics Letters A* **151** (3,4), pp. 107–108.
Peres, A. (1991) 'Two Simple Proofs of the Kochen-Specker Theorem', *Journal of Physics A: Mathematical and General* **24,** L175–L178.
Redhead, M. L. G. (1987) *Incompleteness, Non-Locality, and Realism* (Oxford: Clarendon Press).
Stairs, A. (1983) 'Quantum Logic, Realism, and Value-definiteness', *Philosophy of Science* **50** (4), pp. 578–602.
Walker, R. J. (1950) *Algebraic Curves* (Princeton: Princeton University Press).
Zimba, J. R. (1993) 'Finitary Proofs of Contextuality and Non-locality using Majorana Representation of spin-3/2 States', M.Sc. thesis, Oxford.

228

QUANTUM NON-LOCALITY AND COMPLEX REALITY

The renaissance of general relativity and cosmology: A survey to celebrate the 65th Birthday of Dennis Sciama. Eds G. Ellis, A. Lanza, and J. Miller. Cambridge University Press, 1993

Honouring Dennis Sciama's 65th birthday, this article relates both to the work of Chapters 227 and 291, and to that of Chapters 205, 230, and to the earlier work cited there.

402 Quantum non-locality and complex reality. *The renaissance of general relativity and cosmology: A survey to celebrate the 65th birthday of Dennis Sciama.* Eds G. Ellis, A. Lanza, J. Miller. 1993

Quantum Non-Locality and Complex Reality

ROGER PENROSE

Although I am one of the very few people represented here who was never technically a student of Dennis Sciama's (or a student's student or a student's student's student), I was, on the other hand, very much a student of his in a less formalized sense. He was a close personal friend when I was at Cambridge as a research student, and then a little later as a Research Fellow. Although my Ph.D. topic was in pure mathematics, Dennis took me under his wing, and taught me physics. I recall attending superb lecture courses by Bondi and by Dirac, when I started at Cambridge, which in their different ways were inspirations to me, but it was Dennis Sciama who influenced my development as a physicist far more than any other single individual. Not only did he teach me a great deal of actual physics, but he kept me abreast with everything that was going on and, more importantly, provided the depth of insight and excitement - indeed, *passion* - that made physics and cosmology into such profoundly worthwhile and thrilling pursuits.

I first encountered Dennis at the Kingswood Restaurant, in Cambridge, somewhat before I went up there as a research student, where I was introduced to him by my brother Oliver. I tried to explain to him some geometrical thoughts that I had had in relation to the steady-state universe, as a result of hearing Fred Hoyle's BBC radio talks on the Nature of the Universe, and was impressed by Dennis's keen interest in what even a total unknown such as myself had to say. Already at that time Dennis's sense of excitement about the cosmos came through to me, and it never left me since. In our later discussions, when I was officially installed at Cambridge, Dennis and I had long discussions about relativity, unified field theory, cosmology, quantum theory, particle physics, and especially Mach's principle. I think that apart from gaining a great many necessary insights into existing physical theory from Dennis, especially into general relativity, what I took from those discussions were two things that were particularly important to my own individual (and peculiar?) approaches towards furthering physical understanding. These insights were (i) that some kind of non-locality must be a fundamental feature of the geometry of the actual physical world, and (ii) that there must be an underlying geometric structure to space-time

in which the complex numbers of quantum theory play a role at least on a par with the real numbers of conventional relativity theory.

The importance of non-locality became clear to me as a result of the many discussions that we had concerning Mach's principle. "Suppose", Dennis would often say to me, "that the particles of the distant universe could be eliminated one-by one until only the earth survived. How would a Foucault pendulum then behave?" He would argue, powerfully, that since there would be no distant galaxies to determine the non-rotating frame with respect to which the earth is imagined to be still spinning, it is the earth itself that determines the only frame that is physically present, so no rotation with respect to it is possible and the Foucault pendulum would appear *not* to be deflected. I remember feeling uneasy about this, and we began wondering what would happen if things were taken to even greater extremes, so that there would be only one electron left. What would happen to its spin? No doubt it would have to retain its total spin, so that it would remain a spin 1/2 particle, but what about its spin *direction* It seemed reasonable to me that this concept would become divested of its meaning when there was nothing to compare the electron with, but suppose there were just two electrons. Then each could, in effect, measure its spin in relation to the other. There would be just two possibilities, the total spin of the pair could be 0 ("singlet state") or 1 ("triplet state"). I tried to develop this kind of idea to situations where several particles are involved, in the hope that well-defined directions of spin could arise when large numbers of particles are present. I was eventually able to make such a scheme work (the theory of "spin–networks" cf. Penrose 1971, 1972). One lesson from this study became particularly clear to me: the essential non-locality of physics when the effects of quantum entanglement are considered. It seemed to me that any deep theory of the way that the geometry of the physical world is structured must come seriously to terms with this non-locality.

Passing, now, to the second insight, such a geometric theory, if it were to represent a genuine union between quantum theory and space-time structure would also have to take into account the fact that quantum theory is based on *complex* numbers, rather than the real numbers that underlie the conventional space-time descriptions. Dennis himself, at that time, had been developing a generalization of standard general relativity in which complex numbers played an essential role (Sciama 1958), and we had many discussions together about this and related matters. Partly as a result of such discussions, my own thoughts drove me in the direction of basing space-time geometry on 2-component (complex) spinors - and eventually twistors, where ideas from non-locality also played an important part. (See Penrose and Rindler 1984, 1986; Penrose 1986.)

In the following sections I shall give brief descriptions of two fairly recent ideas that I have had, one concerning quantum non-locality and the other concerning the relationship between twistor theory and general relativity. These will give some idea of my latest thinking on those topics - where Dennis's inspiration has had a profound initial influence on the direction, for good or for bad, of my own work.

1 MAGIC DODECAHEDRA; A NEW EXAMPLE OF QUANTUM NON-LOCALITY

Let me describe a little story - and a puzzle. Imagine that I have recently received a beautifully made regular dodecahedron. It was sent to me by a company of superb credentials, known as "Quintessential Trinkets", who inhabit a planet orbiting the star Betelgeuse. They have also sent another identical dodecahedron to a colleague of mine who lives on a planet orbiting α-Centauri, and his dodecahedron arrived there at roughly the same time as mine did here. Each dodecahedron has a button that can be pressed, on each vertex. My colleague and I are to press individual buttons independently on our respective dodecahedra at some time and in some order that is completely up to our own choosing. Nothing may happen when one of the buttons is pressed, in which case all we do is to proceed to our next choice of button. On the other hand, a bell may ring, when one of the buttons is pressed, accompanied by a magnificent pyrotechnic display which destroys that particular dodecahedron!

Enclosed with each dodecahedron is a list of guaranteed properties that relate what can happen to my dodecahedron and to my colleague's. First, we must be careful to orient our respective dodecahedra in a very precise corresponding way. Detailed instructions are provided by Quintessential Trinkets as to how our dodecahedra are to be aligned, in relation to, say, the centres of the Andromeda galaxy and M-87, etc. The important thing is only that my dodecahedron and that of my colleague must be perfectly aligned with one another. The list of guaranteed properties is, perhaps, fairly long, but all we shall need from them is something quite simple. We must bear in mind that Quintessential Trinkets have been producing things of this nature for a very long time - of the order of a hundred million years, let us say, and they have never been found to be wrong in the properties that they guarantee. The very excellent reputation that they have built up over a million centuries depends upon this, so we can be quite sure that whatever they claim will turn out actually to be true. What is more, there is a stupendous CASH prize for anyone who *does* find them to be wrong!

The guaranteed properties that we shall need concern the following type of sequence of button-pressings. My colleague and I each independently select one one of the vertices of our respective dodecahedra. I shall call these the SELECTED vertices. We do *not* press those particular buttons; but we *do* press, in turn, and in some arbitrary order of our choosing, each of the three buttons that inhabit vertices *adjacent* to the SELECTED one. If the bell rings on one of them, then that stops the operation on that particular dodecahedron, but the bell need not ring at all. We shall require just two properties. These are:

(a) if my colleague and I happen to have chosen diametrically *opposite* vertices as our respective SELECTED ones, then the bell can ring on one of the ones I press (adjacent to my SELECTED one) if and only if the bell rings on the diametrically opposite one of his - irrespective of the particular orders in which either of us may choose to press our respective buttons;

(b) if my colleague and I happen to have chosen exactly *corresponding* vertices (i.e. in the *same* directions out from the centre) as our respective SELECTED ones, then it cannot be the case that the bell does not ring at all for either of us, on the total of six button-presses that we together make.

Now I want to try to deduce something about the rules that my *own* dodecahedron must satisfy independently of what happens on α-Centauri, merely from the fact that Quintessential Trinkets are able to make such strong guarantees without having any idea as to which buttons either I or my colleague are likely to press. The key assumption will be that there is no long-distance "influence" relating my dodecahedron to my colleague's. I shall suppose that our two dodecahedra behave as separate, completely independent objects after they have left the manufacturers.

My deductions are:

(c) each of my own dodecahedron's vertices must be pre-assigned as either a bell-ringer (colour it RED) or as silent (colour it BLUE), where its bell-ringing character is independent of whether it is the first, second, or third of the buttons pressed adjacent to the SELECTED one;

(d) no two next-to-adjacent vertices can be both bell-ringers (i.e. both RED);

(e) no set of six vertices adjacent to a pair of antipodal ones can be all silent (i.e. all BLUE).

We deduce (c) from the fact that my colleague *might* happen to choose, as his SELECTED one, the diametrically opposite one to my own SELECTED one; at least, Quintessential Trinkets will have no way of knowing that he will not. Thus, if one of my three button-presses happens to ring the bell, then we can be sure (by the absence of "influence" assumption) that Quintessential Trinkets must have pre-arranged that particular vertex to a bell-ringer in order to ensure no conflict with (a). Likewise, (d) follows from (a) also. For suppose two next-to-adjacent vertices on my dodecahedron are both bell-ringers. I might have chosen their common neighbour as my SELECTED one and my colleague might have chosen opposite to this as his SELECTED one; moreover my colleague might have elected to press the button first on his dodecahedron which is opposite to the particular bell-ringer on my dodecahedron which is next-to-adjacent to the one that I *actually* first press. Both our bells would have to ring, which contradicts (a). This establishes (d). Finally, (e) follows from (b), together with what we have now established. For suppose that my colleague happens to choose, as his SELECTED one, the vertex *corresponding* to my own SELECTED choice. If none of my three buttons adjacent to this choice is a bell-ringer, then, by (b), one of my colleagues three must be a bell-ringer. It follows from (d) that my own vertex opposite to my colleagues bell-ringer must also be a bell-ringer. This establishes (e).

Now comes the puzzle. Try to colour each of the vertices of a dodecahedron either RED or BLUE consistently according to the rules (d) and (e). You will find that no matter how hard you try you cannot succeed. A better puzzle, therefore, is to provide a *proof* that there is no such colouring. Can it be that, for the first time in a hundred centuries, Quintessential Trinkets, has made a mistake? Recalling the stupendous CASH prize, we eagerly wait the four years, or so, that is required for my colleague's message to arrive, describing what he did and when and whether his own bell rang; but when his message arrives, all hopes of CASH vanish, for Quintessential Trinkets has turned out to be right again!

How do Quintessential Trinkets - or "QT", as they are known for short - actually do it? Of course "QT" really stands for Quantum Theory, and what they have done is to arrange for an atom of spin 3/2 to be suspended at the centre of each of our dodecahedra. These two atoms were produced on Betelgeuse in an initial combined state of total spin 0, and then carefully separated and isolated, so that their combined total spin value remains at 0. Now when either my colleague or I press one of the buttons at a vertex of one of our dodecahedra, a (partial) spin measurement is made, in the direction out from the centre of that particular vertex. More specifically, what is performed is a measurement to determine whether the m-value of the spin, in that direction, is $+1/2$ rather than one of the other three possible values $+3/2$, $-1/2$, or $-3/2$. If the value is indeed found to be $+1/2$, then the bell rings and the pyrotechnic display follows shortly afterwards. If the value is found not to be $+1/2$, then the atom is not observed, and the three unobserved possibilities are carefully

recombined without disturbing their relative phases. Then the process is repeated in another direction, as determined by the next choice of button.

I shall not go into the full details, here, of why (a) and (b) must hold. The essential points are, first, that the $m = +1/2$ states for two next-to-adjacent vertices of a dodecahedron are orthogonal (orthogonal in the Hilbert-space sense, that is, not perpendicular directions); and, second, that the state which is *orthogonal* to a triplet of mutually next-to-adjacent vertex states is itself orthogonal to the state correspondingly related to the antipodal triplet. For details, see Penrose (1992); and for further details, Zimba (1992).

The final conclusion, of course, is that the assumption of no long-distance influence is violated by Quantum Theory and indeed by experiment, according to the conclusions of Aspect and his colleagues (cf. Aspect and Grangier 1986). A virtue of the present scheme (as shared by a number of other, somewhat earlier, recent ideas) is that the puzzling features of this hypothetical experiment are of an entirely yes/no character, with no probabilities arising. For overviews of the earlier work of this nature, and some philosophical discussion, see Redhead (1987) and Brown (1992). Non-locality seems to be here to stay.

2 TWISTORS, MASSLESS FIELDS OF HELICITY 3/2, AND THE EINSTEIN EQUATIONS

It would take too long to provide an overview of twistor theory here, but I should point out some of its salient features. (For detailed accounts, see Huggett and Tod 1985, Penrose and Rindler 1986, and Ward and Wells 1989; for a historical account, see Penrose 1986.) Basically the original idea was to find a way of describing space-time geometry consistently with the above physical requirements: that it should be fundamentally non-local and fundamentally complex. As a first approximation to twistor theory, we may think of twistors as ***light rays*** (null geodesics) in space-time. If the space-time is flat (or at least conformally flat) then the space of such light rays can be described as a real–five–dimensional submanifold $\mathbb{PN}$ of a complex–three–dimensional projective space $\mathbb{PT}$. In fact, $\mathbb{PN}$ divides $\mathbb{PT}$ into two open pieces $\mathbb{PT}^+$ and $\mathbb{PT}^-$. It is also possible to interpret the points of $\mathbb{PT}$ which do *not* lie in $\mathbb{PN}$, by taking into account the fact that physical photons have spin (of fixed absolute value, giving a helicity parallel or antiparallel with the momentum) and *energy*, with the implication that they should be considered as not being exactly localized along any light ray, even though still being treated classically. The points of $\mathbb{PT}$ - $\mathbb{PN}$ ($=\mathbb{PT}^+ \cup \mathbb{PT}^-$) then represent such "photons with energy and helicity". The photons of positive helicity are represented as points of $\mathbb{PT}^+$, and those of negative helicity as points of $\mathbb{PT}^-$. This gives us our second approximation to twistor theory. As third and fourth approximations, we can physically interpret the points of the non-projective space T, which is the complex–four–dimensional vector space from which the projective space PT arises as its space of one-dimensional linear subspaces. The points of T- or *twistors* - can be interpreted, up to a phase multiplier, as providing the energy-momentum structure of any massless particle (third approximation) or as helicity-raising operators (fourth approximation). (There are also *non*-projective versions N, T^+, and T^- of the respective "projective" spaces PN, PT^+, and PT^-.)

The reason that I have used the terminology "approximation" to twistor theory, here, is that I believe that we have not yet found the fully appropriate physical form of the theory - despite the numerous and multifarious applications of twistor ideas that have already come to light. Thus, I am arguing, what has emerged so far constitutes, indeed, merely a set of conceptual approximations to a putative yet–unknown theory. The essential physical difficulty with the twistor theory that presently exists is that it does not fully apply to *general relativity.* Most of the twistor concepts that I have alluded to above depend upon the space-time being flat, or at least conformally flat. Only the first approximation (in terms of light rays) will apply in a general curved space-time, but the trouble here is that the essential *complex* structure disappears, in general, unless the space-time is indeed conformally flat. There are certain twistor concepts that do apply to general curved space-time (local twistors, asymptotic twistors, hypersurface twistors, 2–surface twistors) but these are too limited to provide a global alternative to the conventional space-time description.

There is one line of development that should be mentioned in this context, however. If we allow our "space-time" to become *complex* (with a complex–analytic metric), then its conformal (Weyl) curvature is seen to split into two halves: its *self-dual* and the *anti-self-dual* parts. The self-dual part may be thought of as describing the positive-helicity aspects of the "graviton field" and the anti-self-dual part, its negative-helicity aspects. It is also appropriate to consider "complex vacuum space-times" for which one or the other of these halves of the Weyl tensor is put equal to zero (the Ricci tensor being assumed zero). Such complex 4–manifolds are referred to as *right-flat* (if just the anti-self-dual part of the Weyl curvature survives) or *left-flat* (if just the self-dual part survives). It may be counted as one of the more successful achievements of existing twistor theory that a global complex twistor space T can be constructed for any left-flat complex space-time M and, moreover, the full metric structure of M, including its automatic satisfaction of the Einstein vacuum equations, is fixed in terms of the global structure of the complex manifold T - according to a procedure sometimes known as the "non-linear graviton construction". The (curved) twistor space T can itself be constructed from *free* complex–analytic data (see Penrose 1976). Thus, in a clear sense, twistor theory provides a complete description, in terms of free functions, of the general left-flat (vacuum) space-time.

There has, for a long time, been the tantalizing hope that something of a somewhat similar nature might perhaps be possible for the *general* vacuum equations, when the anti-self-dual condition is removed. Up to this point, attempts towards this goal had been concentrated in one of two directions: that afforded by "ambitwistors" (basically the complexified version of the "light ray" concept of a twistor) and the "googly" approach, where one attempts to describe the *right*-flat spaces using a twistor space rather than a dual twistor space. (The latter would be obtained, trivially, by merely forming the complex conjugate of the non-linear graviton construction.) The hope, with this googly approach, is that the general solution of the Einstein vacuum equations might be obtainable somehow by "adding together", in a suitably non-linear way, twistor spaces for the left-flat and right-flat "parts" of the vacuum space-time, using the original non-linear graviton construction for the former "part" and the googly construction for the latter "part". See Lebrun (1990) for an account of the current status of the ambitwistor approach, and Penrose (1990) for the status of the googly approach. Both approaches encounter fundamental difficulties. There is a basic awkwardness in the ambitwistor approach because, as yet, it has to be phrased in terms of "nth order extensions". Moreover, it is unclear how the construction could be reduced to one involving only free functions. The googly approach, though in principle closer to the original objectives of twistor theory, remains significantly further from a solution than does the ambitwistor approach. The direct googly constructions are as yet somewhat unconvincing as to their "naturalness"; more serious is a lack of any understanding of the non-linear "adding" that would be necessary.

It is for reasons such as these that we may feel encouraged by the existence of a completely fresh approach towards a twistorial description of the Einstein equations. This approach springs from two observations concerning massless fields of helicity $+3/2$. These observations are:

(A) the appropriate field equations governing massless fields of helicity $3/2$ are consistent if and only if the space-time is Ricci-flat;

(B) in Minkowski space-time, the space of conserved charges for massless fields of helicity $3/2$ is twistor space $\mathbb{T}$.

Thus, it would seem that all we need do, in order to construct the elusive twistor space T for a general Ricci-flat space-time M, is to find the space of charges for massless fields of helicity $3/2$ on M. The hope would be that, as was the case for the non-linear graviton construction, the complex space T would be completely equivalent to the space-time M, although related to it only in a non-local way. One might also hope that the general such T could be constructed by the use of free functions, and that twistor theory would thereby establish itself as providing a general procedure for analysing the Einstein vacuum equations. If these aims could be realized, then there would be a powerful case for considering that the physically "correct" twistor concept had at last been found.

Some words of explanation are needed to clarify the meanings of (A) and (B) more fully. In the first place, the "appropriate" field equations referred to in (A) are equations that relate to the (first) *potential* for the field, subject to a gauge freedom. These equations can be taken either in the Dirac (1936) form or (essentially) the Rarita-Schwinger (1941) form. In the Dirac form (using the 2-component spinor notation, cf. Penrose and Rindler 1984), we have a potential $\gamma^{C}_{A'B'}$, symmetric in A' and B', subject to the field equation

$$\nabla^{AA'}\gamma^{C}_{A'B'} = 0,$$

where there is the gauge freedom

$$\gamma^{C}_{A'B'} \rightarrow \gamma^{C}_{A'B'} + \nabla^{C}_{A'}\nu_{B'}$$

for which $\nu_{A'}$ satisfies the neutrino equation

$$\nabla^{AA'}\nu_{A'} = 0.$$

In the Rarita–Schwinger form (in 2–component spinor notation), no symmetry condition is imposed on $\gamma^{C}_{A'B'}$, and we have the pair of field equations

$$\nabla^{A'(A}\gamma^{C)}_{A'B'} = 0, \quad \nabla_{C(C'}\gamma^{C'C}_{A')} = 0,$$

with the same gauge freedom as above, except that $\nu_{A'}$ is not now restricted to satisfy any equation. (Round brackets around indices always denote symmetrization.) The Dirac description is equivalent to that of Rarita–Schwinger, in a special choice of gauge.

The fact that the Ricci-flat condition is necessary and sufficient for the consistent propagation of these equations follows from the work of Buchdahl (1958), Deser and Zumino (1976), Julia (1982), and others.

In *flat* space-time, we can define a field $\varphi_{A'B'C'}$, totally symmetric in its indices, by

$$\varphi_{A'B'C'} = \nabla_{CC'}\gamma^{C}_{A'B'}$$

in the Dirac case, and by

$$\varphi_{A'B'C'} = \nabla_{C(C'}\gamma^{C}_{A')B'}$$

in the Rarita–Schwinger case, which is *gauge invariant*, and satisfies the massless free field equation

$$\nabla^{AA'}\varphi_{A'B'C'} = 0,$$

but these properties fail in a general Ricci-flat space-time. (I am using the term "Ricci-flat" here rather than just "vacuum" to emphasize that the cosmological constant must vanish in this discussion.)

With regard to (B), we must bear in mind that the "charges" being referred to are allowed to be "origin-dependent" quantities, like angular momentum - which would provide examples of the charges for massless fields in the case of spin 2. Since we are now considering Minkowski space-time, we can use the gauge-invariant description of the helicity $+3/2$ field. To see how twistors arise as the charges for such fields, we introduce a spinor $\mu^{A'}$ subject to the "dual twistor equation"

$$\nabla^{A(A'}\mu^{B')} = 0,$$

which has exactly four independent solutions. The solution space may be identified as the *dual* $\mathbb{T}^*$ of the twistor space $\mathbb{T}$(cf. Penrose and Rindler 1986). Now this equation implies the important property that if we form the quantity

$$\chi_{A'B'} = \varphi_{A'B'C'}\mu^{C'},$$

we find that it satisfies the massless free field equation

$$\nabla^{AA'}\chi_{A'B'} = 0,$$

but now for helicity 1 (so $\mu^{A'}$ acts as a helicity–lowering operator). Thus, $\chi_{A'B'}$ is the spinor form of a free self-dual *Maxwell* field. If we consider a world–tube in Minkowski space-time which is surrounded by a $\chi_{A'B'}$ satisfying this equation, then we can perform the appropriate Gauss integral

$$\int \chi_{A'B'}dx^{A'}_{A} \wedge dx^{AB'}$$

over any 2–surface surrounding the tube, in order to obtain its charge. But this charge is a (complex–)linear expression in $\mu^{C'}$, so the "charge" for the original field must be an element of the *dual* space to the space of the quantities $\mu^{C'}$, i.e. it must belong to *twistor* space $\mathbb{T}$.

This establishes (B), but if we try to generalize this procedure to a curved Ricci-flat space-time, we run into all sorts of problems. (See Penrose 1992, for a taste of what is involved.) The root of the difficulty seems to be the fact that most concepts of "conserved charge" lead to objects that constitute a vector space, which is not what we expect, in general here. Even in the anti-self-dual case, the twistor space T is a particular kind of complex manifold, not a complex vector space. The way around this difficulty seems to be to generalize the linear object described by $\gamma^{C}_{A'B'}$ to something non-linear, most likely to a connection of some kind. The most promising idea in this direction seems to be to take this to be a connection on an SL(3,C)–bundle over the space-time. This is still "work in progress" and it remains to be seen where it will all lead.

REFERENCES

Aspect, A. and Grangier, P. (1986) "Experiments on Einstein–Podolsky–Rosen–type correlations with pairs of visible photons", in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Brown, H.R. (1992) "Bell's other theorem and its connection with nonlocality. Part I.", preprint; lecture at Cesena conference.

Buchdahl, H.A. (1958) "On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field", *Nuovo Cim.*, **10**, 96-103.

Deser, S. and Zumino, B. (1976) "Consistent supergravity", *Phys. Lett.*, **62B**, 335-7.

Dirac, P.A.M. (1936) "Relativistic wave equations", *Proc. Roy. Soc. (Lond.)*, **A155**, 447-59.

Huggett, S.A. and Tod, K.P. (1985) "An Introduction to Twistor Theory", London Math. Soc. student texts (L.M.S. publ.)

Julia, B. (1982) "Système linéaire associé aux équations d'Einstein", *Comptes. Rendus. Acad. Sci. Paris, Sér. II*, **295**, 113-6.

Lebrun, C.R. (1990) "Twistors, ambitwistors and conformal gravity", in *Twistors in Mathematics and Physics*, eds T.N.Bailey and R.J.Baston (Cambridge Univ. Press, Cambridge), 71-86.

Penrose, R. (1971) "Angular momentum: an approach to combinatorial space-time", in *Quantum theory and Beyond*, ed. Ted Bastin (Cambridge University Press, Cambridge).

Penrose, R. (1972) "On the nature of quantum gravity", in *Magic without Magic*, ed. J.R.Klauder (Freeman, San Francisco).

Penrose, R. (1976) "Non-Linear gravitons and curved twistor theory", *Gen. Rel. Grav.*, **7**, 31-52.

Penrose, R. (1986) "On the origins of twistor theory", in *Gravitation and Geometry*, (I. Robinson festschrift volume), eds. W. Rindler and A. Trautman (Bibliopolis, Naples).

Penrose, R. (1990) "Twistor theory after 25 years - its physical status and prospects", in *Twistors in Mathematical Physics*, eds T.N.Bailey and R.J. Baston (Cambridge Univ. Press, Cambridge).

Penrose, R. (1992a) "Twistors as spin 3/2 charges", in *Gravitation and Modern Cosmology* (P.G.Bergmann's 75th birthday vol.) eds. A. Zichichi, N. de Sabbata, and N. Sanchez (Plenum Press, New York).

Penrose, R. (1992b) "On Bell non-locality without probabilities: some curious geometry", in CERN volume in honour of J.S Bell, ed. J. Ellis.

Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge).

Penrose, R. and Rindler, W. (1986) *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry*, (Cambridge University Press, Cambridge.

Rarita, W and Schwinger, J. (1941) "On the theory of particles with half-integer spin", *Phys. Rev.*, **60**, 61- .

Readhead, M.L.G. (1987) *Incompleteness, Nonlocality, and Realism* (Clarendon Press, Oxford).

Sciama, D.W. (1958), *Nuovo Cim.*, **8**, 417.

Ward, R.S. and Wells, R.O. Jr (1989) *Twistor Geometry and Field Theory* (Cambridge University Press, Cambridge).

Zimba, J. (1992) Oxford preprint.

229

ON THE COHOMOLOGY OF IMPOSSIBLE FIGURES

The visual mind: Art and mathematics, 27–9, Leonardo Book Ser., Cambridge, 1993 [French: 'La cohomologie des figures impossibles.' Dual French-English text. *Structural Topology* No. 17 (1991), 11–16]

This explains the cohomology referred to in Chapter 168. In later work (not covered by the period of these Collected Works) I have considered that this kind of picture may be useful in understanding how the twistor representation of the wavefunctions of massless particles exhibits non-local features in relation to quantum measurement. (See Penrose, R. (2005) The twistor approach to space-time structures. In: *100 Years of Relativity; Space-time Structure: Einstein and Beyond*, Ed. A. Ashtekar, World Scientific, Singapore.)

On the cohomology of impossible figures. *The visual mind: Art and mathematics*, 27–9, 1993
[French: 'La cohomologie des figures impossibles.' Dual French-English text. *Structural Topology* No. 17 (1991), 11–16]

CHAPTER

5

On the Cohomology of Impossible Figures

Roger Penrose

ABSTRACT

The close relationship between certain types of impossible figure and the mathematical idea of cohomology is explained in relation to the tribar and to another type of impossible figure related to the Necker cube.

In a recent article [1], presented in honour of M. C. Escher, I hinted at a relationship between cohomology and certain types of impossible figure. It is the purpose of this note to explain this relationship more fully.

I shall be concerned with the concept of *first* cohomology group

$$H^1(Q, G); \qquad (1)$$

the basic meaning of this concept should emerge during the course of the discussion. Here Q is some (non-simply connected) region of the plane—which I shall take to contain the 'support' (i.e. the region of the plane where the drawing occurs) of some impossible figure—and G is a (normally Abelian) group, which I shall refer to as the *ambiguity group* of the figure. (For those readers not familiar with the mathematical concept of a group, it may be taken that G is just some set of numbers closed under multiplication and division. Thus if a and b belong to G, then so do ab and a/b.) To fix ideas, let us consider two examples. The first is the tribar, illustrated in Fig. 1. Here, Q can be taken to be, say, the region of the plane (paper) on which the tribar is actually drawn, or else some slightly larger region such as the annular region depicted in Fig. 2. In the second example, illustrated in Fig. 3, I have drawn a version of impossible figure that I introduced in my earlier article.

Consider first the tribar. We may regard the region Q as being pasted together from three smaller regions Q_1, Q_2, Q_3, as indicated in Fig. 4. There are overlapping pars of Q_1, Q_2, Q_3, which are to be pasted together.

The drawing, on each of Q_1, Q_2, Q_3, is a perfectly consistent rendering of a three-dimensional structure that is unambiguous in its natural interpretation—except for the essential ambiguity present in all pictures: one does not know the *distance* away from the observer's eye that the object being depicted is supposed to be situated (Fig. 5). Of

---

Roger Penrose (mathematician), Mathematical Institute, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, United Kingdom.

This article was originally published in *Structural Topology* 17 (1991) pp. 11–16. Reprinted with permission.

Fig. 1. An impossible figure, the tribar, drawn in perspective.

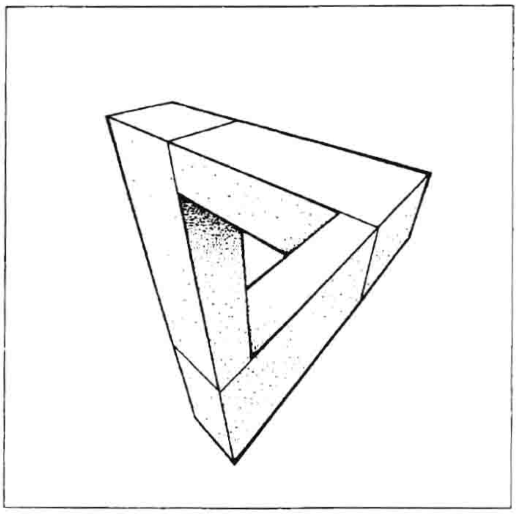
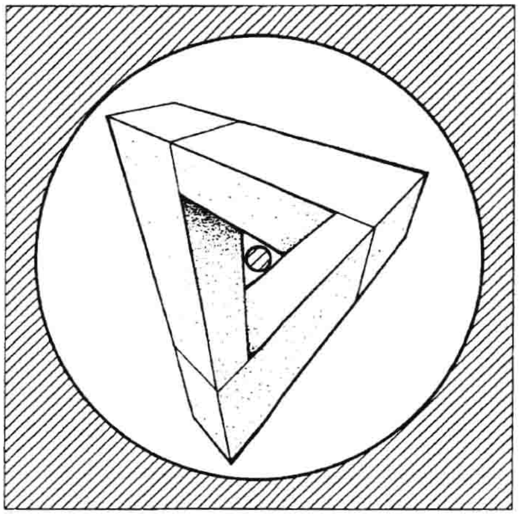


Fig. 2. The tribar, drawn on an annular region of the plane.



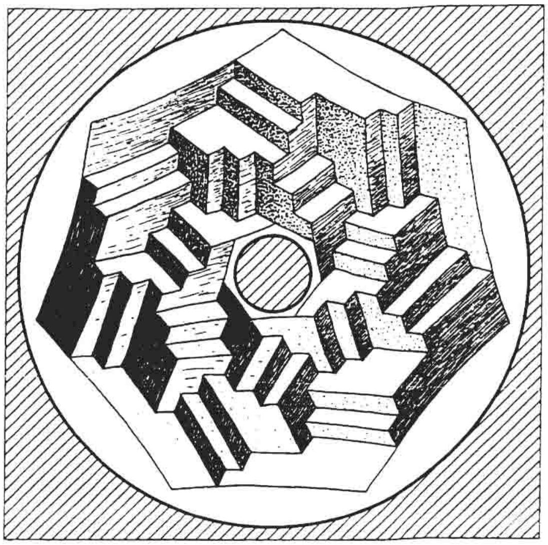


Fig. 3. A more subtle impossible figure with local Z_2 ambiguity, drawn on an annular region of the plane.

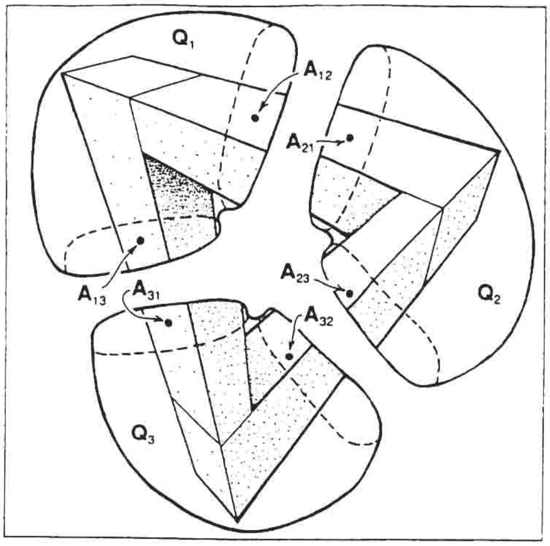


Fig. 4. The tribar shown pieced together out of overlapping smaller drawings, each of which depicts a possible structure.

course there are always other ambiguities, such as the fact that the picture could be depicting a *picture* of another picture, for example, rather than a three-dimensional object (a feature that Escher often put to paradoxical use, e.g. in his lithograph *Drawing Hands* and woodcut *Three Spheres I*). I am excluding this and other possible ambiguities here by my use of the phrase 'natural interpretation'. This distance can be described in terms of positive real numbers d, the set of all possible real numbers being denoted by $\mathbf{R}^+$. I am thinking of $\mathbf{R}^+$ as a multiplicative (Abelian) group, so in this case we have the ambiguity group $G = \mathbf{R}^+$. Let us see how this comes about.

Consider the portion of the figure drawn in region Q_1, and fix a point A_{12} on this portion where it overlaps with Q_2, and a point A_{13} on it where it overlaps with Q_3. Let A_{21} be that point of the figure, as drawn on Q_2, that is to be matched with the point A_{12} on Q_1, and similarly let A_{31} be the point on Q_3 that is to be matched with A_{13}. Finally fix a point A_{23} on the part of the figure on Q_2 that is to be pasted on Q_3, and the corresponding point A_{32} on Q_3 that is to be matched with it. See Fig. 4 for the entire arrangement of points.

Let us suppose that there is an actual three-dimensional object O_1, which the drawing on Q_1 depicts and, similarly, actual objects O_2 and O_3, which the drawings on Q_2 and Q_3 depict (see Fig. 5). The point on O_1 that is depicted by A_{12} may not be the same distance from the observer's eye E as the corresponding point on O_2, depicted by A_{21}. Let the ratio of these distances be d_{12}, and similarly for other pairs of matched points. Thus we have

$$d_{ij} = \frac{\text{distance from E to point on } O_i \text{ depicted by } A_{ij}}{\text{distance from E to point on } O_j \text{ depicted by } A_{ji}} \qquad (2)$$

We note first that d_{ij} does not actually depend on the particular matched pair of points A_{ij}, A_{ji}, which are chosen on the overlap between Q_i and Q_j. We get the same d_{ij} whichever such matching pair we choose. This d_{ij} represents the factor that we must move out by when we pass from O_j to O_i at the region of overlap.

Note also that

$$d_{ij} = 1/d_{ji} \qquad (3)$$

and that if we change our minds about the object O_i that is being depicted in Q_i (i.e. if we change its chosen distance from the observer's eye) then the pair (d_{ij}, d_{ik}) is replaced according to

$$(d_{ij}, d_{ik}) \rightarrow (\lambda d_{ij}, \lambda d_{ik}), \qquad (4)$$

for some positive number λ.

If, instead of the tribar, we had had some drawing of a figure that could be consistently realized in three-dimensional space, then we could have moved the objects O_1, O_2 and O_3 in and out until they all came together as one consistent structure. This amounts to the fact that by rescalings of the above type we can reduce the three ratios d_{12}, d_{23} and d_{31} simultaneously to 1. Another way of saying this is that there exist three (positive) numbers q_1, q_2, q_3 such that

$$d_{ij} = q_i/q_j \qquad (5)$$

for each different i, j. In the terminology of cohomology theory, the collection $\{d_{ij}\}$ is, in the general case, referred to as a *cocycle*. If (5) holds, the cocycle is called a *coboundary*. The replacement (4) provides the *coboundary freedom*, and we regard cocycles as *equivalent* if they can be converted to one another under this freedom. Under this equivalence, we obtain the *cohomology group elements*, i.e. the elements of

$$H^1(Q, \mathbf{R}^+). \qquad (6)$$

The coboundaries provide the *unit* element of (6), and we see from the above discussion that the test for whether or not the figure depicted in Q is 'impossible' is whether or not the resulting element of (6) is indeed the unit element.

I have been discussing impossible figures of the kind that I described earlier [2] as 'pure', i.e. for which the only local ambiguity in the figure is the *distance* from the observer's eye

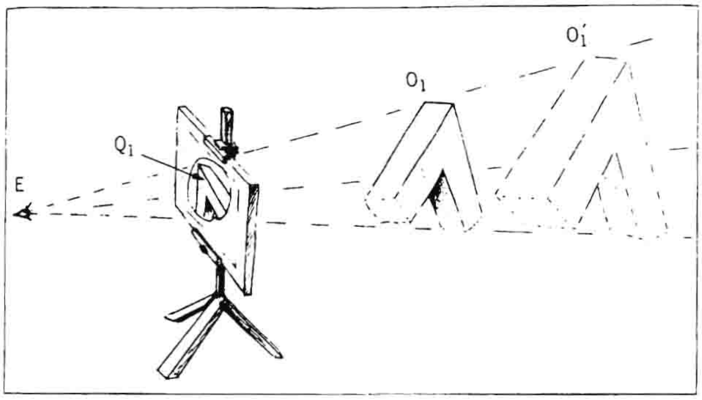


Fig. 5. There is a local R^+ ambiguity in any plane drawing as to the distance from the observer's eye to the object depicted.

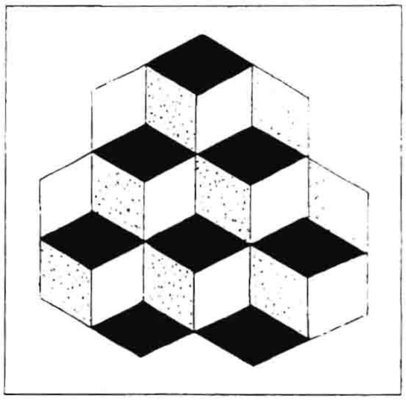


Fig. 6. Necker cubes, with Z_2 ambiguity.

of the object being depicted. Often there are other ambiguities of relevance. For the type of impossible figure depicted in Fig. 3, the relevant ambiguity is that of the 'Necker cube', see Fig. 6. Here the ambiguity is just a twofold one, and we can use the numbers +1 and −1 in place of the distance ratios d_{ij} defined in (2), where +1 means that the depicted three-dimensional object O_i *agrees* with O_j where the drawings overlap, and −1 means that the objects *disagree.* The discussion proceeds exactly as before, except that d_{ij}, λ and q_i now all belong to $\mathbf{Z}_2$ (the multiplicative group consisting of +1 and −1 alone), and the cohomology group element we obtain belongs to

$$H^1(Q, \mathbf{Z}_2). \tag{7}$$

If we cut Fig. 3 into three pieces analogous to those of Fig. 4 and follow the corresponding procedure through, we indeed find an element of (7) that is *not* the unit element, whereas if Fig. 3 has been drawn 'consistently' (e.g. with a hexagon—or, indeed, an octagon—at the centre, rather than a heptagon), then the unit element would have been obtained. I leave the detailed verification of these facts to the interested reader.

More complicated figures with 'multiple impossibilities' [3] can also be analyzed in this way, but for this we should require a more complete description of what a (Cech) cohomology group actually is. In general, the figure would need to be divided up into more than three pieces, but the essential idea is the same as before [4]. I believe that considerations such as these may open up intriguing possibilities for further exotic types of impossible figure. I hope to be able to consider such matters at a later date.

References

1. R. Penrose, "Escher and the Visual Representation of Mathematical Ideas", in H. S. M. Coxeter, M. Emmer, R. Penrose and M. L. Teuber, eds., *M. C. Escher: Art and Science* (Amsterdam: North Holland, 1986) pp. 143–147.

2. See Penrose [1].

3. See, for example, L. S. Penrose and R. Penrose, "Impossible Objects: A Special Type of Visual Illusion", *Brit. J. Psych.* 49 (1958) pp. 31–33.

4. For further information, see P. Griffiths and J. Harris, *Principles of Algebraic Geometry* (New York: Wiley, 1978) p. 34.

230
SPIN 3/2 FIELDS AND LOCAL TWISTORS

Twistor Newsletter **37** (1994). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer, 353–9. Chapman and Hall/CRC, London, 2001. With L. J. Mason

This represents another attack on the issues raised by Chapters 200, 202, 203, and 205.

420 Spin 3/2 fields and local twistors. *Twistor Newsletter* **37** (1994). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer, 353–9. 2001. With L. J. Mason

§III.4.25 **Spin $\frac{3}{2}$ fields and local twistors** *by L.J. Mason and R. Penrose* (𝕋ℕ 37, April 1994)

Introduction. In §III.4.19 and §III.4.20 one of us (RP) proposed that helicity $\frac{3}{2}$ fields might provide a suitable vehicle for the definition of a twistor in vacuum space-times. On the one hand, in flat space-time, twistors emerge as the charges of helicity $\frac{3}{2}$-fields, and on the other, the vacuum equations are the consistency condition for the existence of such fields (in potential form, e.g., that given by Rarita-Schwinger) so that they can exist only when the field equations are satisfied.

There are still fundamental obstacles. The difficulties of defining charges in curved space were discussed in Penrose (1992). However, there are intriguing issues that are already present in a flat background. We have to deal with the description of the field as a potential modulo gauge rather than as a field. (The helicity $\frac{3}{2}$ 'field' associated with a Rarita-Schwinger potential is not gauge invariant in curved space.) Thus, a deeper examination of the R-S potential, its gauge freedom, and its relation to twistor theory should prove fruitful.

A neat form of the R-S equations can be given in terms of spinor-indexed forms. Let $\sigma_{A'}$ be a spinor-indexed 1-form and d be the exterior derivative extended to act on spinor-indexed quantities. Then the gauge freedom is $\sigma_{A'} \to \sigma_{A'} + d\nu_{A'}$ where $\nu_{A'}$ is a spinor-indexed function, and the R-S equations are

$$dx^{AA'} \wedge d\sigma_{A'} = 0. \qquad (1.1)$$

One can see directly that the gauge transformations will be consistent with the field equations iff

$$dx^{AA'} \wedge d^2\nu_{A'} = dx^{AA'} \wedge R^{B'}_{A'}\nu_{A'} = 0$$

for all $\nu_{A'}$. This follows iff the Ricci tensor vanishes (this is effectively the same calculation as that required to show that the Sparling 3-form is closed when the vacuum equations are satisfied—see Penrose & Rindler 1986, Vol. 2 Chapter 6). There are 8 equations for 8 unknowns, and two free functions worth of gauge freedom, so we require two relations between the field equations for the system to be consistent (this can be seen by removing two of the unknowns using the gauge freedom to set, for example, $V \lrcorner \sigma_{A'} = 0$ for some vector field V so that there are 2 more equations than unknowns). The relations follow by taking the exterior derivative of (1.1). This vanishes identically iff $dx^{AA'} \wedge R^{B'}_{A'} = 0$ so the equations are consistent in vacuum.

In flat space, the field is given by

$$d\sigma_{A'} = \psi_{A'B'C'}dx^{BB'} \wedge dx^{C'}_{B}$$

and is gauge invariant, but in curved space a gauge transformation adds $\Psi_{A'B'C'D'}\nu^{D'}$ to $\psi_{A'B'C'}$, making it ill-defined.

Since the helicity $\frac{3}{2}$ field is only directly described in potential form in curved space, in Penrose (1992), a description of charges was given that only involves the potential. We shall consider the general case of a 1-form potential γ for the dual field ${}^*F$ associated to a 2-form field F perhaps taking values in some flat vector bundle. Introduce a covering of the relevant region of space-time M by open sets $\{U_i\}$. The (dual) field on this region is equivalent to a collection of potentials γ_i on each U_i such that $d\gamma_i = {}^*F$ so that on $U_i \cap U_j$, γ_i and γ_j differ by a gauge transformation (the collection $\{\gamma_i\}$ are taken to be defined modulo global gauge transformations). The charge associated with $\{\gamma_i\}$ takes values in the space of *gauge freedom of the second kind.* For electromagnetism, this formulation follows from the de-Rham sequence for potentials of the dual field

$$0 \to \mathbb{C} \to \mathcal{E}(M) \to \Omega^1(M) \to \Omega^2(M) \to \ldots$$

where $\mathcal{E}(M)$ is the space of smooth functions on M and the second map is the injection of constant functions and the third and subsequent maps are the standard exterior derivative. The charge q of an electromagnetic field is obtained by integrating the (closed) 2-form $^*F$ over some given 2-surface, $\mathcal{S}$. The closed 2-form $^*F$ is equivalent to a collection $\{\gamma_i\}$ with $\gamma_i \in \Omega^1(U_i)$ and $d(\gamma_i-\gamma_j)=0$ on $U_i \cap U_j$ defined up to $\gamma_i \to \gamma_i + df_i$ for some $f_i \in C^\infty(M)$. By diagram chasing (or direct integration), one can see that q lies in the $\mathbb{C}$ of the above sequence playing the role of a *gauge transformation of the second kind*, which is to say a gauge transformation that leaves the *potential* invariant. This follows by defining f_{ij} by $\gamma_i - \gamma_j = df_{ij}$. It then follows, with $q_{ijk} = f_{ij} + f_{jk} - f_{ik}$, that $dq_{ijk} = 0$ so q_{ijk} is a Čech cocycle with constant coefficients and one can see that its evaluation on $\mathcal{S}$ is the charge (up to the usual 4π, etc.).

This point of view for helicity $\frac{3}{2}$ fields led to a new difficulty even in flat space: if one only has the Rarita-Schwinger field, then the analogue of the above sequence is now

$$0 \to \mathbb{S}_{A'} \to \mathcal{E}_{A'} \to \Omega^1_{A'} \to \Omega^2_{A'} \to \dots$$

where the third and fourth maps are the covariant exterior derivative. The problem is that the gauge transformations of the second kind are just constant spinors $\mathbb{S}_{A'}$ whereas we were hoping to obtain a whole twistor. This spinor is the secondary part of the charge of the corresponding helicity $\frac{3}{2}$ field. In Penrose (1992, 1995) it was noted that if one wishes to encode the primary part of the charge, one requires the next potential down the potential chain. A proposal for an exact sequence involving this second potential in the R-S case was also given in Penrose (1995).

The purpose of this note is to give an improved version of this sequence, with unrestricted gauge freedom, and relate it to local twistors.

Local twistors and the helicity $\frac{3}{2}$ equations. In Woodhouse (1985) and Mason (1990) it was noted that the potential chains for zero-rest-mass fields have a natural formulation in terms of local twistors. Perhaps the simplest formulation of the potential chain for the helicity $\frac{3}{2}$ equation is as follows

$$\psi_{A'B'C'} = \nabla^A_{A'}\sigma_{B'C'A}, \qquad \sigma_{AA'B'} = \nabla^B_{A'}\rho_{B'AB}, \qquad \rho_{ABA'} = \nabla^C_{A'}\phi_{ABC}$$

where each potential is a symmetric spinor, and is subject to field equations and gauge freedom (cf. Penrose and Rindler 1986 Vol. 2, section 6.4).

A slightly different formulation of the potential chain, without its first and last parts, can be encoded in a local-twistor-valued 1-form with $n-2$ symmetric indices for helicity n. This is easily understood from the positive homogeneity twistor-function description:

for helicity $\frac{3}{2}$ one takes a dual twistor function R of homogeneity one and then evaluates $R^\alpha = \partial R/\partial W_\alpha$ as a self-dual Maxwell Field with a local twistor index. Denote this field by $\mathcal{R}^\alpha$. (Since, at this stage, we are working in conformally flat space-time, the bundle of local twistors is flat and so we can just evaluate such an object in a covariantly constant frame component by component and then transform to the standard one.) Then $\mathcal{R}^\alpha$ is defined modulo gauge transformations, $\mathcal{R}^\alpha \to \mathcal{R}^\alpha + DZ^\alpha$ where Z^α is an arbitrary function with values in local twistors and D denotes the exterior derivative extended to act on quantities with a local twistor index using the local twistor connection. We also have the self-duality equation $(D\mathcal{R}^\alpha)^- = 0$ where the superscript '$-$' on a 2-form denotes its A.S.D. part. As a consequence of the fact that R^α is of the special form $\partial R/\partial W_\alpha$, $D\mathcal{R}^\alpha$ also has vanishing primary part. This follows from the fact that the 'field' integral formula for the primary part is the expression

$$\oint \frac{\partial^3 R}{\partial \tilde{\pi}_A \partial \tilde{\omega}^{A'} \partial \tilde{\omega}^{B'}} \tilde{\pi}^C d\tilde{\pi}_C.$$

However, this is of the form of the integral of $\partial f_{-1}/\partial \tilde{\pi}_A \tilde{\pi}^B d\tilde{\pi}_B$ where f_{-1} has homogeneity -1 (and has indices) but this is an exact form being proportional to $d(\tilde{\pi}^A f_{-1})$ so that its integral vanishes.

This last condition can be written $X_{\alpha\beta} D\mathcal{R}^\beta = 0$ where

$$X_{\alpha\beta} = \begin{pmatrix} \varepsilon_{AB} & 0 \\ 0 & 0 \end{pmatrix} \tag{2.2}$$

is the 'position twistor,' i.e., the canonical section (up to scale) of $\mathbb{T}_{[\alpha\beta]}$ over Minkowski space given by expressing points of Minkowski space as elements of the projectivisation of $\mathbb{T}_{[\alpha\beta]}$ via the Klein correspondence.

The primary and secondary parts are the potentials for the helicity $\frac{3}{2}$ field as follows

$$\mathcal{R}^\alpha := \begin{pmatrix} \rho^A \\ \sigma_{A'} \end{pmatrix} := \begin{pmatrix} \rho^A_b dx^b \\ \sigma_{A'b} dx^b \end{pmatrix}.$$

The equations $(D\mathcal{R}^\alpha)^- = 0$ and $X_{\alpha\beta} D\mathcal{R}^\beta = 0$ become

$$d\rho^A + i dx^{AA'} \wedge \sigma_{A'} = 0 \quad \text{and} \quad (d\sigma_{A'})^- = 0.$$

The first equation can be seen to imply the second by first taking the covariant exterior derivative of the first (in flat space still) to obtain

$$dx^{AA'} \wedge d\sigma_{A'} = 0,$$

the form of the R-S equation given in the introduction, and then writing out this system in full. The helicity $\frac{3}{2}$ field is then the secondary part of $D\mathcal{R}^{\alpha}$.

Thus, locally at least, $\mathcal{R}^{\alpha}$ modulo DZ^{α} satisfying $X_{\alpha\beta}D\mathcal{R}^{\alpha}=0$ is equivalent to a helicity $\frac{3}{2}$ field and is the $\mathcal{P}$-transform of $\partial R/\partial W_{\alpha}$.

This formulation of the potential for a helicity $\frac{3}{2}$ field is sufficient to encode the full charge of the field in the context of gauge freedom of the second kind since we now have the exact sequence

$$0\rightarrow\mathbb{T}^{\alpha}\rightarrow\mathcal{E}^{\alpha}\rightarrow\Omega^{1\alpha}\rightarrow\Omega^{2\alpha}\rightarrow\dots$$

where the second map is the injection of the covariantly constant local twistors into general local twistor fields, the third and fourth maps are the exterior derivative extended to act on local twistors using the local twistor connection. Thus the gauge freedom of the second kind is given precisely by constant twistors. (A sequence of this nature was given in Penrose 1995, but with a restricted gauge freedom.)

Generalization to curved space The above equations do not work as they stand in curved space as, when $D^2\neq 0$, the equation $X_{\alpha\beta}D\mathcal{R}^{\beta}=0$ is no longer compatible with the gauge freedom. However, a weaker equation is compatible with the gauge freedom in vacuum, namely,

$$DX_{\alpha\beta}\wedge D\mathcal{R}^{\beta}=0. \tag{3.3}$$

There is a subtlety here as $X_{\alpha\beta}$ is only defined up to scale by the conformal metric. The scale is fixed by choosing a particular conformal factor and representing $X_{\alpha\beta}$ by equation (2.2). Equivalently, a choice of conformal scale leads to an infinity twistor

$$I_{\alpha\beta}=\begin{pmatrix}0 & 0\\ 0 & \varepsilon^{A'B'}\end{pmatrix}$$

which can be used to normalize the position twistor by means of the relation

$$X_{\alpha\beta}I^{\alpha\beta}=2.$$

Equation (3.3) is only conformally invariant when $X_{\alpha\beta}D\mathcal{R}^{\alpha}$ vanishes also. This additional equation can be consistent only in flat space, however.

The compatibility of the gauge freedom with equation (3.3) follows iff we have the relation $DX_{\alpha\beta}\wedge\mathcal{K}^{\beta}_{\gamma}=0$ where $\mathcal{K}^{\alpha}_{\beta}$ is the curvature of the local twistor connection and $X_{\alpha\beta}$ is normalized as above by a choice of conformal factor. This relation holds iff

$\nabla^{A}_{A'}\Psi_{ABCD} = 0$ and thus requires the Einstein equations. As before, there are as many equations as unknowns (in this case 16 rather than 8) but the gauge freedom can be used to set some (in this case 4) of the unknowns to zero leading to an overdetermined system. However, if the relation $DX_{\alpha\beta} \wedge \mathcal{K}^{\beta}_{\gamma} = 0$ holds, the covariant exterior derivative of equation (3.3) vanishes identically leading to four relations between the 16 equations so that they are not overdetermined in vacuum (or more generally when the trace-free part of the Ricci tensor vanishes). (Conversely, if $\nabla^{A}_{A'}\Psi_{ABCD} \neq 0$ the equations are inconsistent.)

If we write out this system in terms of the primary and secondary parts of $\mathcal{R}^{\alpha}$, we obtain the equations

$$dx^{B'}_{A} \wedge d\rho^{A} + i dx^{B'}_{A} \wedge dx^{AA'} \wedge \sigma_{A'} = 0 \quad \text{and} \quad dx^{AA'} \wedge \sigma_{A'} = 0.$$

In flat space we see that we acquire an extra helicity-$\frac{3}{2}$ field, since

$$D\mathcal{R}^{\alpha} = \begin{pmatrix} \tilde{\psi}^{A}_{BC} dx^{BB'} \wedge dx^{C}_{B'} \\ \psi_{A'B'C'} dx^{BB'} \wedge dx^{C'}_{B} \end{pmatrix}$$

with the ψ's totally symmetric and $\nabla^{A}_{A'}\tilde{\psi}_{ABC} = 0 = \nabla^{A'}_{A}\psi_{A'B'C'}$. Conversely given such ψ's, we can write down the twistor indexed 2-form as above which is closed and so can be written as $D\mathcal{R}^{\alpha}$ for some $\mathcal{R}^{\alpha}$ and automatically satisfies $DX_{\alpha\beta} \wedge D\mathcal{R}^{\alpha} = 0$.

However, the gauge freedom of the second kind is still just the constant local twistors, since the fields sit in the same exact sequence as the single helicity $\frac{3}{2}$ fields earlier.

Where do we go from here?

Connections with other aspects of twistor theory: There are some intriguing connections with other aspects of twistor theory. Firstly, the connections between twistor theory and integrable systems highlight the role played by linear systems in twistor correspondences and so one might hope that the above linear system might lead to some kind of twistor construction. See Mason (1995) for an exploration of this line of reasoning. Secondly, the identities that guarantee the consistency of the R-S equations are precisely those that lead to the characterization of the vacuum equations by means of the Sparling 3-form. This suggests that a (local) twistorial analogue of the Sparling 3-form should be taken to be $DZ^{\alpha} \wedge DZ^{\beta} \wedge DX_{\alpha\beta}$. It is certainly closed iff the vacuum Bianchi identities are satisfied. This expression (with some modification to eliminate the terms that contain the Ricci tensor explicitly) also has an interpretation as a Hamiltonian that generates a translation along the vector field $\pi^{A'}\omega^{A}$ together with a spin-frame rotation generated by $\pi_{A'}\pi^{B'}$. These ideas should extend to a local twistor generalization of the connections between the Sparling 3-form, the canonical formalism and quasi-local mass described in Mason and Frauendiener (1990).

Towards a definition of twistors in vacuum space-times

The aforementioned considerations serve to clarify the role of twistors and their relationship to the R-S equations in flat space, by exhibiting them as gauge quantities of the second kind. However, the full generalization of these ideas to Ricci-flat curved space remains elusive. So far, these considerations suffer from being 'too linear', since a twistor space without a vector space structure ought eventually to arise. One promising route to achieving this would be to examine a role for twistors as providing a 'charge' in the *active* sense rather than passive sense, similarly to the way in which the electric charge features in the electromagnetic connection $\nabla_a - ieA_a$. Combining the active with the passive roles for a twistor might provide a route to the required nonlinearity. Work is in progress.

Thanks to J. Frauendiener and G.A.J. Sparling for comments and discussion.

References

Mason, L.J. (1990) Local twistors and the Penrose transform for homogeneous bundles, in *Further Advances in Twistor Theory*, §I.2.17 (see also §I.2.18), Vol. I, eds. L.J. Mason & L.P. Hughston, Pitman Research Notes in Mathematics No. **231**, Longman.

Mason, L.J. (1995) Generalized twistor correspondences, d-bar problems and the KP equations, in *Twistor Theory*, ed. S. Huggett, Lecture Notes in Pure and Applied Mathematics, **169**, Marcel Dekker.

Mason, L.J. & Frauendiener, J. (1990) The Sparling 3-form, Ashtekar variables and quasi-local mass, in *Twistors in Mathematics and Physics*, eds. T.N. Bailey and R.J. Baston, LMS lecture notes series **156**, Cambridge University Press, Cambridge.

Penrose, R. & Rindler, W. (1984, 1986) *Spinors and Space-Time*, Vols. 1, 2, Cambridge University Press, Cambridge.

Penrose, R. (1992) Twistors as spin $\frac{3}{2}$ charges, in *Gravitation and Modern Cosmology,* eds. A. Zichichi, N. de Sabbata, and N. Sánchez, Plenum Press, New York.

Penrose, R. (1995) Twistors and the Einstein equations, in *Twistor Theory*, ed. S. Huggett, Lecture Notes in Pure and Applied Mathematics, **169**, 145-57, Marcel Dekker.

Woodhouse, N.M.J. (1985) Real methods in twistor theory, Class. & Quant. Grav. **2** 257-291.

231

MATHEMATICAL INTELLIGENCE

What is intelligence? 107–36. Darwin College Lectures, Cambridge University Press, 1994

It is my hope that in this *Darwin Lecture* I was able to express, in simple direct terms, the essential mathematical arguments underlying my belief that we are going to have to move far beyond the commonly believed computer model of mental activity, if we are ever going to comprehend what is involved in the mental activity of 'intelligent understanding'.

5

Mathematical intelligence

ROGER PENROSE

What is mathematical intelligence? Is there anything that is essentially different about the way that we reason mathematically, from the way in which we think generally? Is mathematical intelligence different from any other kind of intelligence?

I feel certain that there is no fundamental difference between mathematical and other kinds of thinking. It is true that many people find it difficult to cope with the abstract type of thinking that is needed for mathematics, whilst finding comparatively little difficulty with the equally convoluted judgements that are involved in day-to-day relationships with other human beings. Some kinds of thinking come easily to certain people, whereas other kinds come more easily to others. But I do not think that there is any essential difference – or that there is more difference between mathematical thinking and, say, planning a holiday, than there is between the latter activity and understanding a music-hall joke. Human mathematical intelligence is just one particular form of human intelligence and understanding. It is more extreme than most of these other forms in the abstract, impersonal, and universal nature of the concepts that are involved, and in the rigour of its criteria for establishing truth. But mathematical thinking is in no way removed from other qualities that are important ingredients in our general ability for intelligent comprehension, such as intuition, common-sense judgement, and the appreciation of beauty.

What, after all, is intelligence? What is thinking? There is a prevalent viewpoint in current philosophising that holds that whatever it is that in detail constitutes the physical activity underlying our thought processes, it cannot, in effect, be other than the carrying out of some vastly complicated calculation. The relevant actions of our brains, so it is argued, are simply to provide our bodies with a very effective control system – a control system that could in principle be effected by a computer, if only one knew enough of the details of those computational procedures that the brain actually carries out. One might well imagine that, in accordance with this view, such an underlying computational basis to our thinking ought to be most manifest with mathematical thinking. For is not mathematics a computational activity *par excellence*? Indeed, it is not! It is one of my purposes here to emphasise that there is a great deal of what is essential in mathematical thinking that is not of a computational character. Indeed, it turns out that it is possible actually to demonstrate that there is something in our mathematical understanding – in our insights as to mathematical truth – that eludes any computational description whatever. It is the very precision and the universal character of mathematical argument that allows such a demonstration to be possible. But the conclusion is in no way restricted to an intelligence that relates merely to mathematical thinking. As I have argued, there is nothing essential that separates mathematical from other types of thinking, so our demonstration that mathematical understanding is something that cannot be simulated in a computational way can be thought of, also, as a demonstration that understanding itself – one of the most essential ingredients of genuine intelligence – is something that lies beyond any kind of purely computational activity.

MATHEMATICAL VISUALISATION

What are we doing when we conjure up in our minds the image of some mathematical structure? Are we performing some internal calculation, like those that lead to the impressive computer graphic displays with which we are now so familiar? Perhaps our brains are

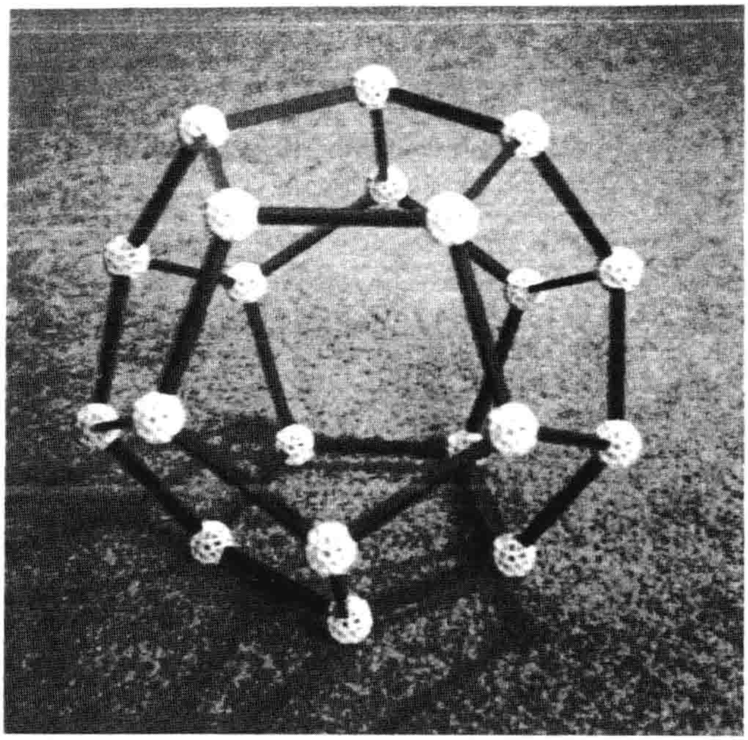


Figure 1 A regular dodecahedron

acting out something like the computational procedures that give rise to what is called 'virtual reality', whereby an entire three-dimensional seemingly consistent structure, such as a non-existent building, can be visually presented to a human subject, through the agency of a pair of special stereoscopic goggles. The detailed scene that each eye perceives is the result of a complicated calculation performed in 'real time' so that the structure appears to remain consistent no matter how the subject turns his head or moves his body. Are we doing something similar when, in our 'mind's eyes' we conjure up some consistent mental image of a three-dimensional object, whether real or entirely imagined?

I shall argue that we are actually doing something very different from this. Let us consider an example. Figure 1 is a photograph of a regular dodecahedron. With some effort, it may be possible for us to rotate this image to a different orientation. In fact, we may feel that we have some conception of the object as an actual three-dimensional structure rather than as something that needs a particular vantage

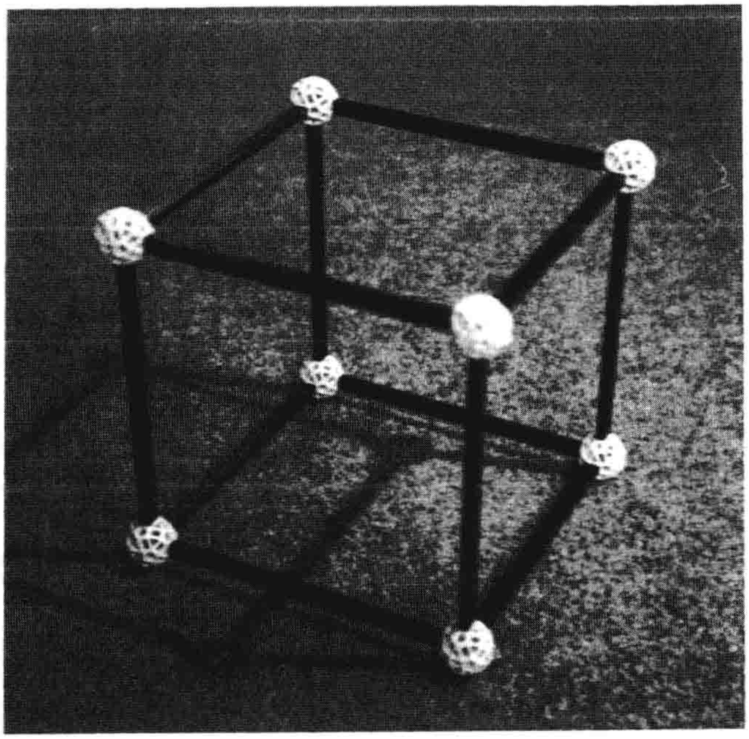


Figure 2 A cube

point from which to view it. Many people would find considerable difficulty in visualising an entire dodecahedron, but the cube depicted in the photograph of Figure 2 is a good deal easier. It is not that hard to transfer the flat image on the page to a 'solid' three-dimensional imagined structure. This may seem to be similar to what is involved with computer graphic displays. In Figure 3, I have provided a sequence of computer images of a regular dodecahedron viewed from successively slightly different vantage points, so that the

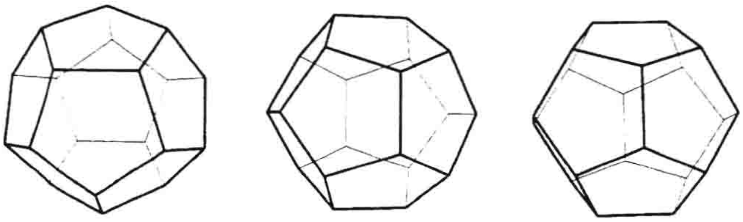


Figure 3 Computer pictures of a dodecahedron, from a gradually moving vantage point

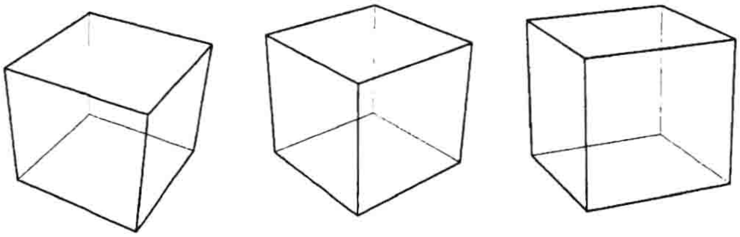


Figure 4 Computer pictures of a cube, from a gradually moving vantage point

dodecahedron appears to rotate just as it would if we slowly move around it. In Figure 4, I have done the same thing with a cube. In each case, there is, inside the computer, a stored representation of the dodecahedron, or cube, that does not change, but the chosen vantage point is gradually altered. Might not our own visual images be something like such stored computer representations?

I think it is unlikely – and to support this contention, let me indicate some fairly clear-cut distinctions. In the first place, the computer displays are far more accurate than anything that can be at all easily achieved by human imagination. Of course, it might well be argued that we are simply being very inefficient and inaccurate in our visualisations, as compared with a modern computer. Indeed, it would not be hard to introduce inaccuracy into our computer simulations, so that they fall to the level of accuracy that would be relevant to any particular human individual. If it were just our inaccuracy that distinguishes our own acts of visualisation from the outputs of computers then my argument would certainly be a very weak one. But visualisation carries with it strong elements of understanding, and it is actual understanding that the computer simulations lack.

To illustrate this point, consider Figure 5. Here I have added some lines to the photograph of Figure 1 to show that a cube can be found inside the dodecahedron, its eight vertices coinciding exactly with a selection of eight from the twenty vertices of the dodecahedron. It is not hard to see that this selection of eight vertices indeed gives us an exact cube. Symmetry considerations alone will tell us this; each face must clearly be an exact rectangle, at least, and the rectangle's sides must indeed be equal since each is a 'diagonal' of one of the equal

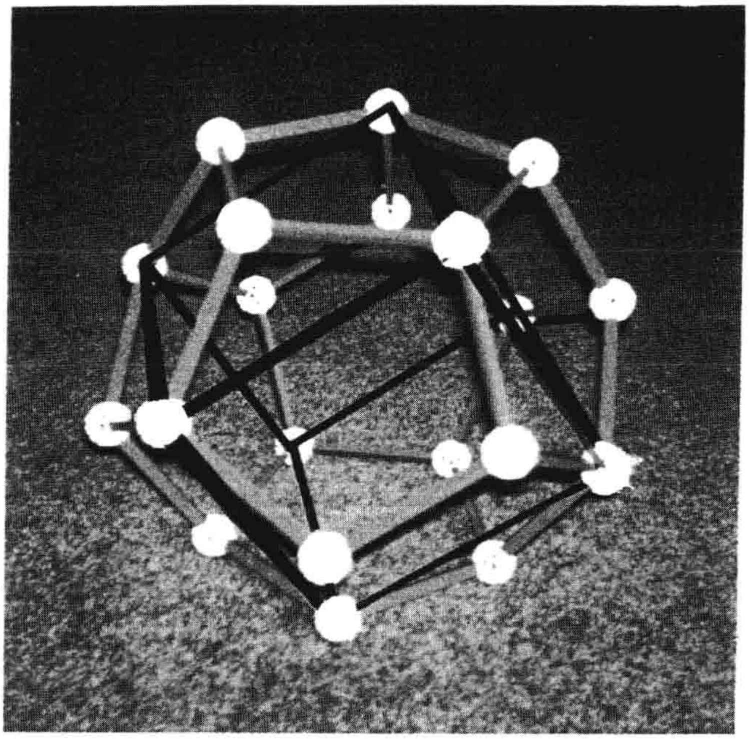


Figure 5 Dodecahedron with additional lines to show that a cube sits within it, sharing eight of its vertices

regular-pentagonal faces of the dodecahedron. Can the computer simulation 'see' this fact? This position would be hard to maintain, I feel, but we can at least test it by asking whether or not two of the edges of the computer's proposed 'cube' are equal. (This, in itself, is far from sufficient, of course, but at least it is a start.) When I tried this on two edges, the computer came out with two ten-digit decimal expressions that differed in the final place. When I asked it whether these two numbers were equal, it asserted that indeed they were. (Apparently, the program allows for some round-off error.)

This is hardly a convincing demonstration that the computer in any sense 'knew' that there is an absolutely exact cube in the regular dodecahedron. Rather, it seems to me, it establishes the contrary conclusion, that all this particular computer 'visualisation' can do is come up with approximations, albeit approximations with nine or ten figures of accuracy. It has no way of reaching the exact conclusion that our own visualisations – and accompanying understandings – are capable of: that indeed our proposed cube is geometrically precise.

In fact, being wise after the event, a computer programmer might replace the particular way in which the dodecahedron is stored in the machine by another one whereby exact information about distances and angles could be retrieved on demand. It would then be possible for the machine to give a correctly affirmative answer to our question about whether the suggested cube is indeed an exact one. Actually, being wise after the event, there would be an even easier way: the computer could be instructed simply to answer 'yes' to this particular question about the cube! The trouble with this, of course, is that the computer itself could in no way be said to possess any mathematical understanding of the exactness of the cube. It would simply be parrotting the information that its programmer had provided it with, and no-one would argue that any understanding of the exactness lay with the computer rather than with its human programmer.

One might try to do better than this, of course, and perhaps equip the computer with the complete system of axioms for three-dimensional Euclidean geometry. It could then try to ascertain whether a given statement, such as the exactness of the cube referred to above, could be deduced from these axioms. In this way it could, in principle at least, provide the correct answers to many geometrical problems. Of course, it might still be questioned whether what the computer does bears any relation to what a human mathematician does when understanding that a geometrical statement is actually true. That human understanding has to do with a belief in the validity of those intuitions – based to a good extent on symmetry considerations – underlying the very choice of the axioms themselves. The issue is a somewhat delicate one for there are valid geometric axiom systems that are distinct from those of Euclid. Indeed, when I later present powerful arguments in support of the thesis that our insights and understandings are not things that can be reduced to computation, it will be necessary to turn away from geometry, and to address the issue of computation directly, where it will be our intuitions concerning the natural numbers 0, 1, 2, 3, 4, 5, . . . , rather than geometrical forms, that will be the subject of our deliberations.

But before turning to such matters, let me give credit to what might be called one of the early 'success stories' of Artificial Intelligence. In the early 1960s, H. L. Gelernter programmed a computer to derive propositions in Euclidean plane geometry from the axioms with which it had been initially provided. When the computer came up with its proof that if a triangle has two equal sides, then the angles opposite to those sides are also equal, Gelernter was startled. For the computer's proof had been unknown to him, and it was considerably simpler than that given by Euclid. The computer's argument was this (see Figure 6): since AB=AC, the triangles ABC and ACB are congruent (side-side-side); therefore <ABC equals <ACB, QED! In fact this argument was not new. (It was given in the fourth century AD by Pappus.) But it was undoubtedly a striking fact that a computer could come up with something so elegant and unexpected.

In this example, the computer's success arose because its 'blind' rule-following stopped it from being distracted from the seeming absurdity of its own argument. No doubt there are many other situations in which human mathematicians have been similarly distracted from seeing arguments that they should have seen. However, this particular example is one for which the chain of reasoning is very short, and it is not hard to find it by means of a mindless search. When the derivations from axioms are long and complicated, as they tend to be with mathematical arguments of considerable sophistication, then the 'mindless search' method becomes hopelessly inefficient.

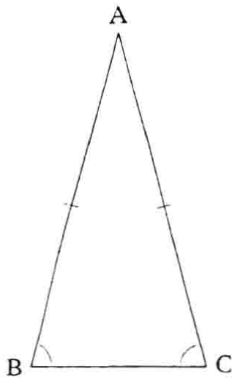


Figure 6 An isosceles triangle; the angles at B and C are equal – as proved by Gelernter's computer program, since ABC and ACB are congruent

Insight and understanding are necessary ingredients, if the search is to become manageable.

PROOF BY GEOMETRICAL INSIGHT 1: FAREY FRACTIONS

Choose a reasonable-sized natural number n (say $n=9$) and write down, in order of size, all the fractions from 0 to 1, expressed in their lowest terms, whose denominators do not exceed n (here with $n=9$)

0/1, 1/9, 1/8, 1/7, 1/6, 1/5, 2/9, 1/4, 2/7, 1/3, 3/8, 2/5, 3/7, 4/9, 1/2, 5/9, 4/7, 3/5, 5/8, 2/3, 5/7, 3/4, 7/9, 4/5, 5/6, 6/7, 7/8, 8/9, 1/1

Such an array is referred to as a sequence of *Farey fractions*. Let me point out one remarkable property that such a sequence possesses. If we take the difference between any pair of consecutive fractions in the list, then we find that the numerator is always 1. For example

$$2/5-3/8 = (16-15)/(5\times 8) = \mathbf{1}/40$$
$$4/7-5/9 = (36-35)/(7\times 9) = \mathbf{1}/63$$

How do we know that this must always be the case? We must find a way of convincing ourselves that if a/b and c/d are consecutive fractions in the list, then

$$ad - bc = 1$$

(since $a/b-c/d=(ad-bc)/bd$). To do this, we can use a geometrical argument. Let us imagine each fraction a/b to be plotted as the point (a,b) on a graph, where we also take the points representing successive fractions to be joined by straight-line segments. Figure 7 gives an accurate representation of all this, for $n=9$, but the figure is rather crowded and it will be much clearer if we use the inaccurate representation of part of the sequence that is illustrated in Figure 8. (It is a striking fact about mathematical diagrams, and the attendant visualisation of abstract concepts – which is indeed what is going on here – that an inaccurate image, if it is inaccurate in an appropriate way, may often give more of an insight as to what is going on than

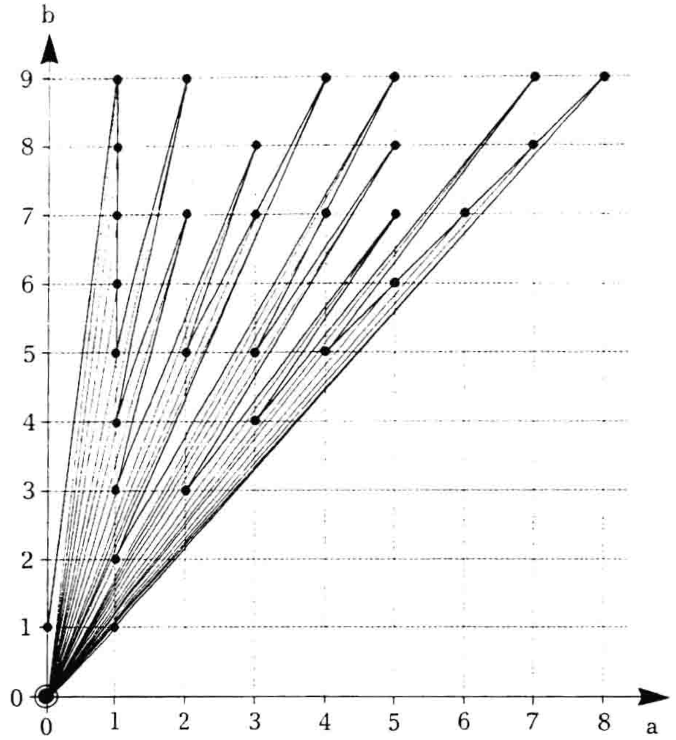


Figure 7 Farey fractions illustrated geometrically

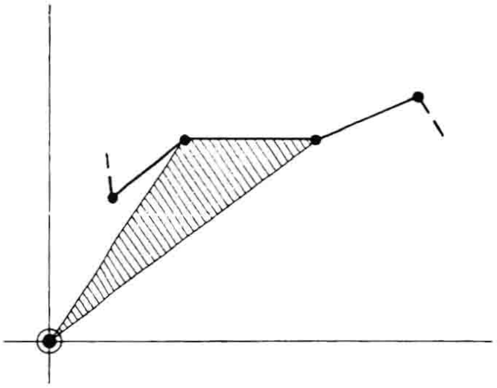


Figure 8 The area of the shaded triangle is $\frac{1}{2}$ unit

an accurate one!) We wish to show that $ad-bc=1$. It is a well-known formula of coordinate geometry that the *area* $\triangle$ of the triangle whose vertices are (0,0), (a,b) and (c,d) is given by

$$\triangle = (ad - bc)/2$$

so what we must establish is that this area is precisely 1/2 (Figure 9).

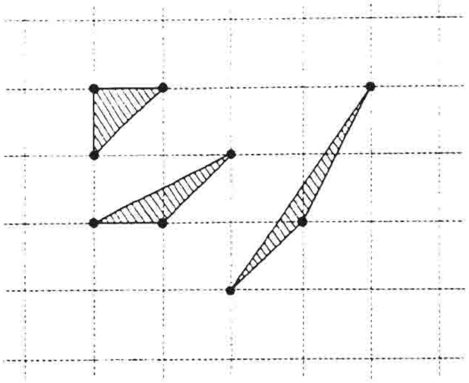


Figure 9 Triangles whose vertices are on lattice points and which contain no other lattice points, either internally or on an edge, are each of area = ½ unit

In fact, it is a theorem (essentially a special case of one due to Minkowski) that if a triangle's vertices all lie at integer lattice points (the points (x,y) where x and y are integers) and if it has the property that it contains no other lattice point either in its interior or on its edges, then it must have an area equal to 1/2. Thus, what we need to show is that our triangle indeed has this property. This is actually not hard to see from the fact that the two fractions a/b and c/d are both in their lowest terms (so that there are no further lattice points on the sides of the triangle out from the origin (0,0)), and from the fact that all fractions with denominators no greater than n have been included – a fact that would be contradicted if our triangle constructed from two successive fractions a/b and c/d contained any other lattice point.

It remains to establish the aforementioned theorem. One way of doing this (Figure 10) is to consider a sequence of transformations whereby the triangle is successively moved without changing its area until it becomes just half of a lattice square. Each transformation consists of taking one of the vertices and moving it parallel to the opposite side, until it reaches another lattice point that lies closer to the other two vertices than before. I shall not bother with the full details of this here, but I hope that the rough idea is clear.

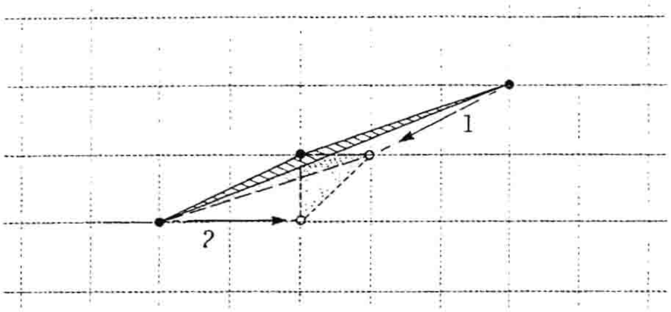


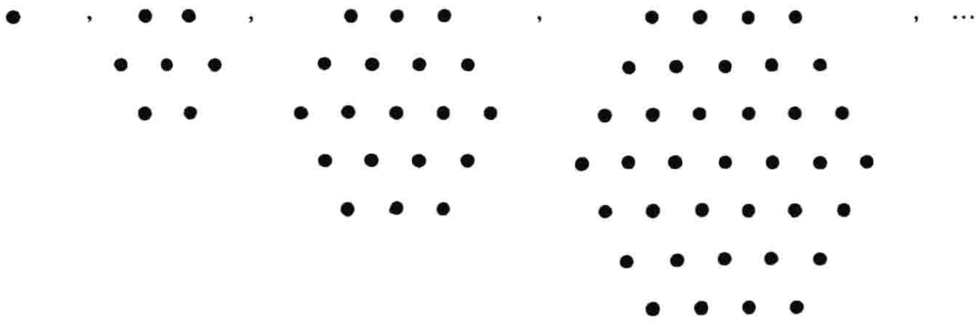
Figure 10 The fact that the area is ½ unit can be proved by successively displacing a vertex to another lattice point closer to the opposite side, in a direction parallel to that side, until a diagonally bisected square is obtained

PROOF BY GEOMETRICAL INSIGHT 2: HEXAGONAL NUMBERS

My next example involves what are called hexagonal numbers

1, 7, 19, 37, 61, 91, 127, . . .

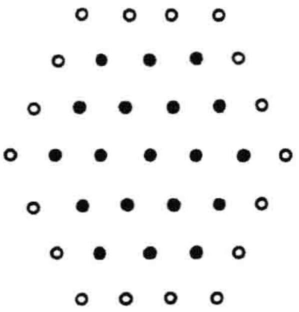
namely the numbers that can be arranged as regular hexagonal arrays (excluding the vacuous array):



These numbers are obtained, starting from 1, by adding successive multiples of 6

6, 12, 18, 24, 30, 36, . . .

as we see from the fact that each hexagonal number can be obtained from the one before it by adding a hexagonal ring around its border:



For the number of spots in this ring is a multiple of 6, the multiplier increasing by 1 each time, as the hexagon gets larger.

Now let us add together the hexagonal numbers successively, up to a certain point, starting with 1. What do we find?

$$1=1,\ 1+7=8,\ 1+7+19=27,\ 1+7+19+37=64,\ 1+7+19+37+61=125$$

The numbers 1, 8, 27, 64, 125 are all *cubes*. A cube is a number multiplied by itself three times

$$1=1^3=1\times1\times1,\ 8=2^3=2\times2\times2,\ 27=3^3=3\times3\times3,\ 64=4^3=4\times4\times4,$$
$$125=5^3=5\times5\times5,\ \ldots$$

Is this a general property of hexagonal numbers? Let's try the next case. We indeed find

$$1+7+19+37+61+91 = 216 = 6\times6\times6 = 6^3$$

I am going to try to convince you that this is always true. First of all, a cube is called a cube because it is a number that can be represented as a cubic array of spots

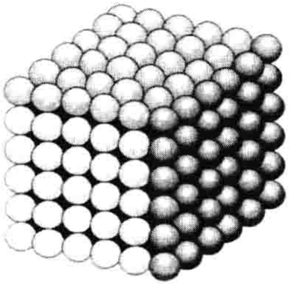


Figure 11 A cubical array of spheres

I want you to try to think of such an array as built up successively, starting at one corner and then adding a succession of three-faced arrangements each consisting of back wall, side wall and ceiling, as depicted thus:

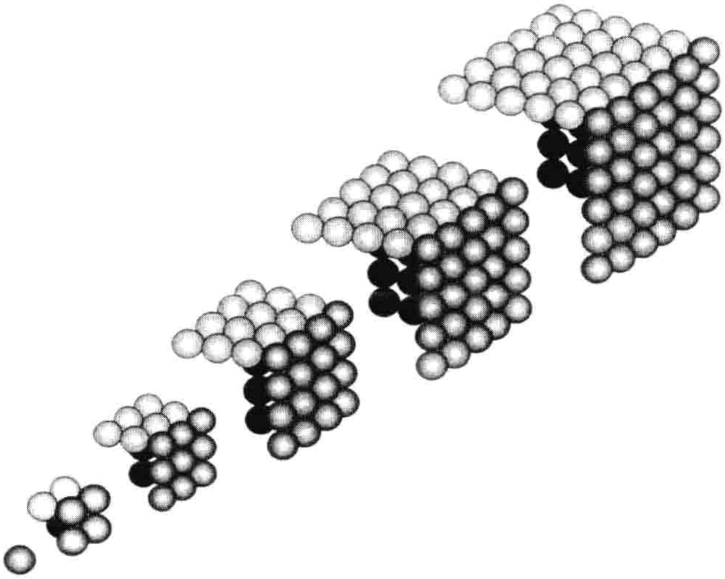


Figure 12 The cubic array is separated into a succession of layers consisting of side wall, back wall and ceiling, each of which is viewed from a long distance off

Now view this three-faced arrangement from a long way out, along the direction of the corner common to all three faces. What do we see? A hexagon:

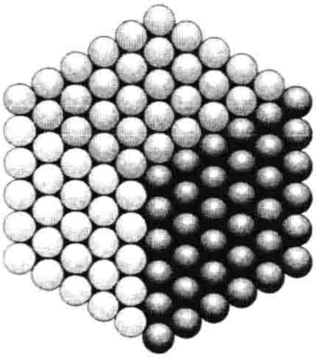


Figure 13 Each layers appear as a hexagonal arrangement – a hexagonal number

The spots that constitute these hexagons, successively increasing in size, when taken together, correspond to the spots that constitute the entire cube. This establishes the fact that adding together successive hexagonal numbers, starting with 1, will always give a perfect cube.

PROOF BY GEOMETRICAL INSIGHT 3: RULES OF ARITHMETIC

This last example again shows the power of mathematical (geometrical) visualisation. Let us try something else; this time a good deal more elementary. How do we know that, for two natural numbers a and b, we always have

$$a + b = b + a\,?$$

For this, all we need to do is visualise a collection of things (imagined to be 'a' in number) to which we add another collection of things (imagined to be 'b' in number). The total number of things altogether is clearly the same whichever order we add them in, so the required result follows.

This example is very trivial, and it is perhaps not quite 'geometrical' in the ordinary sense, but I think that it represents a genuine mathematical insight from an act of imagination. One might try to argue that this insight is really just an aspect of our 'experience' of the persistence of objects in the world. For example, apples and oranges do not just disappear when placed in a box, nor do they magically appear within the box. But there is more to it than this. None of us has ever directly experienced precisely 88990012345 objects or 60606999931 objects, yet we would have no doubt that indeed

$$88990012345 + 60606999931 = 60606999931 + 88990012345.$$

Our experiences with apples and oranges merely act as a guide towards our mathematical insights. Those insights are genuine abstractions that are valid methods of reasoning about abstract mathematical objects. (As always, in mathematics, one must be

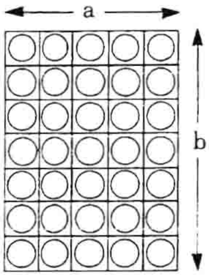


Figure 14 A rectangular array of a columns of b objects gives the same total number as would b columns of a objects – by rotation of the figure

extremely careful, however. There are kinds of *infinite* numbers for which the rule $a+b=b+a$ fails – but that's another story!)

Let's try another rule of elementary arithmetic:

$$a \times b = b \times \mathrm{a}$$

This tells us that if we are to imagine taking 'a' collections of objects, where each collection contains 'b' objects, then the total number would be unaltered if we did the same thing but with 'a' and 'b' interchanged. Stated this way, it is not really obvious; but if we imagine our collections to be arranged as a succession of 'a' columns, where each column contains 'b' objects, then the symmetry becomes obvious (Figure 14). We can imagine rotating the resulting rectangular array through a right angle to achieve this. Even easier than this (if we have mental difficulty in visualising the rotation) is simply to read our array off the other way around rather than mentally rotating it, that is as row-by-row rather than column-by-column. This works just as well for, say,

$$97666000011 \times 777708999 = 777708999 \times 97666000011$$

as for 5×7=7×5, even though we cannot precisely visualise collections of things that represent these actual individual numbers.

How about the associative law

$$(a \times b) \times c = a \times (b \times c)$$

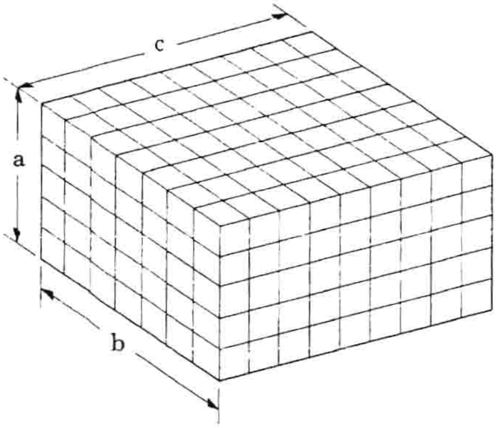


Figure 15 A three-dimensional rectangular array shows that $(a \times b) \times c = a \times (b \times c)$

We can see this one by imagining a three-dimensional rectangular array which we read off in two different ways: see Figure 15. Again this works just as well for, say

$$(97666000011 \times 777708999) \times 83383302222$$
$$= 97666000011 \times (777708999 \times 83383302222)$$

as for $(5 \times 7) \times 9 = 5 \times (7 \times 9)$, even though these large numbers cannot be individually visualised precisely.

An important thing to note about these 'visualisations' is that they are not really pictures of actual things in space, but of something much more abstract. It really does not matter whether 'actual' space has the necessary accurately Euclidean structure, or that it extends outwards far enough so the particular numbers that we wish to represent can be realised in terms of actual objects within that space. (For example, if we wished to represent the number $10^{10^{1000}}$, which is a perfectly good number, still subject to the same algebraic laws as the numbers that we can directly visualise, then we could not do so within the 'observable universe' with actual physical objects.) Nevertheless, our simple visualisations are indeed sufficient to provide the necessary insights that can convince us that the algebraic properties that we have been considering are actually true of all natural numbers, no matter how large they may be.

Now, let us consider the relation

$$((a\times b)\times c)\times d = a\times(b\times(c\times d))$$

Can we be sure that this one is universally true? Now we use a different method. It is not much good to use a four-dimensional array, which is what would be needed if we were to try to visualise the relation directly. We have no direct experience of four spatial dimensions, so our geometrical intuitions are not of much immediate use in this case. Instead, we can turn to algebra and deduce this last relation from what we have established previously. First

$$((a\times b)\times c)\times d = (a\times b)\times(c\times d)$$

by the earlier relation (but with $a \times b$, c, d, in place of a, b, c, respectively); then

$$(a\times b)\times(c\times d) = a\times(b\times(c\times d))$$

again by this earlier relation (but now just with $c\times d$ in place of c). The result follows by combining these two.

Algebra provides a very useful means of replacing our direct insights by calculational procedures. We do not now have to think of what our expressions actually 'mean'. We can now just calculate! In fact this is somewhat overstating the case I want to make. The effective use of algebra often requires a good deal of understanding, subtlety, and even artistry. But in many ways, it is the power of a good calculus (like ordinary algebra or, indeed, ordinary arithmetical notation) that, to a considerable degree, it enables understanding to be temporarily suspended and replaced by blind calculation. It is in this facility for blind calculation, rather than understanding, that computers can far exceed the capabilities of even the most effective of human experts.

CAN UNDERSTANDING ALSO BE REDUCED TO BLIND CALCULATION?

What about the human faculty of understanding? Can we be sure that it is not itself some kind of calculational activity? Presumably anyone who believes that genuine Artificial Intelligence is possible must also believe that the quality of understanding can be artificially simulated – for understanding is surely an essential part of genuine intelligence. The present meaning of the term Artificial Intelligence (or 'AI') is that it is something according to which the artificial simulation is indeed performed calculationally, by which I mean by the use of electronic computers. Although it is possible to imagine that, in the future, some means of 'artificial simulation' which is different from anything that can be achieved calculationally might be introduced, I shall stick to the standard 'calculational' meaning of the term here. Accordingly, those who believe that genuine AI is possible must believe that the quality of understanding can indeed be simulated calculationally.

Before giving my main reason for disbelieving this possibility, I should be a little clearer about what I mean by the term 'calculational'. In effect, as I indicated earlier, I mean anything that can be performed by a modern general-purpose computer. This is perhaps not very precise as a mathematical definition. What I really mean, technically, is anything that can be performed by a Turing machine. However, I appreciate that many readers will not know precisely what the term 'Turing machine' means, so it is easier just to refer to a modern general-purpose computer, where we require this 'computer' to be a mathematically idealised concept. The idealisations that we require are that the computer never makes mistakes, that it can continue indefinitely without ever wearing out, and that it has an unlimited storage capacity. If you think of this last idealisation as being a little unreasonable, just imagine that it is always possible to add more storage capacity (that is 'memory') to the computer whenever it runs out.

An important point to make here is that actions of things like 'parallel computers' and (artificial) 'neural networks' (or 'connection machines'), that we hear quite a lot about these days, are all included in what I mean by 'calculations'. For some types of problem, a computer designed according to what is called a 'parallel architecture' may be much more efficient or much faster than an ordinary serial computer, but there is no difference between the two in principle. Likewise, systems like neural networks – referred to as 'bottom-up' systems – which improve upon their performance by a 'learning' process aimed at optimising the quality of their output, are also calculational. A bottom-up action contrasts with the standard 'top-down' calculational procedures in that the latter operate according to an algorithm that is known to work correctly for the class of problem that it is concerned with, whereas with bottom-up action no such algorithm is given beforehand and, instead, a means is initially provided whereby the system is to improve its performance as it gains experience. This is still a calculational procedure (and therefore still an 'algorithm'), however, because the very means whereby the system is to improve its performance is itself given by a set of calculational rules. From the point of view of the present discussion, the essential difference between a bottom-up and a top-down system is that the former is only an approximate means of obtaining the required answers even though it may sometimes be a very effective one.

A simple test for deciding whether a system is calculational is to ask: can it be run on an ordinary computer? If it can be then the system is indeed a calculational one. In fact, as far as I am aware, most of the (artificial) neural network systems that have been constructed to date are actually, as they stand, simulations run on ordinary computers – so that there is no question but that they must be calculational! (Perhaps it is intended eventually to construct some special electronic hardware, on which the neural network system would be run much more efficiently, but that makes no difference to the fact that such systems are always calculational in nature.)

I now wish to present an argument that effectively demonstrates that mathematical understanding is not a calculational activity. The argument is based on a form of the famous theorem of Kurt Gödel, that he proved in 1930, but where I also call upon some later ideas introduced mainly by Alan Turing in about 1935. I shall not need any of the technical details of Gödel's argument, and although the argument may be found to be somewhat confusing, it does not use any difficult mathematics.

The type of calculation that we shall be concerned with will be an operation that can be performed on a natural number. The action of some calculation C on a natural number n is written C(n). We may think of C as being given by a computer program, where after feeding the program into our computer, we then supply the computer with the number n, which the computer operates on to produce the answer. (Technically, C may be thought of as a Turing machine, and C(n) is the action of that Turing machine on the natural number n.) I shall not be very concerned here with the actual result of the calculation, but mainly whether or not it ever eventually stops.

Let us consider some examples. One possible calculation would be to form the square n^2 of the natural number n. This particular calculation encounters no problem about eventually stopping, since the square of any given natural number can certainly be formed in a finite time. (Recall that there is to be no limit on the computer's storage space.) More subtle is the following example of a calculation, which depends on the given natural number n.

> *Find the smallest natural number that is not the sum of* n *squares.*

Our calculation would proceed, trying the natural numbers in turn: 0, 1, 2, 3, 4, 5, . . . ; until it finds one that is not the sum of n squares. To get the idea, let us first see how this works for $n=2$. We start with 0, and find that 0 is indeed the sum of two squares, namely $0=0^2+0^2$. We must move on to 1, and find that although 0^2+0^2 doesn't work, we do find that indeed $1=0^2+1^2$. Thus, we must move on to 2, finding that although 0^2+0^2, or 0^2+1^2, or 0^2+2^2 do not work, we indeed have $2=1^2+1^2$. Moving on to 3, we find that none of 0^2+0^2, 0^2+1^2, 0^2+2^2, 0^2+3^2, 1^2+1^2, 1^2+2^2, 1^2+3^2, 2^2+2^2, 2^2+3^2, or 3^2+3^2, will work (cutting the calculation off when the number to be squared reaches the number to be summed to – though we could be more efficient, cutting things off earlier). Thus we find 3 as the smallest number that is not the sum of two squares. We could now try this all over again with $n=3$, finding, in this case, that the number 7 is the smallest that is not the sum of three squares. We can also go back and test the case $n=1$, finding that 2 is the smallest number not the sum of one square (and examining the logic involved in the case $n=0$, we find that 1 is the smallest number not the sum of zero squares).

Now let us consider $n=4$. Our calculation proceeds, finding

$$0=0^2+0^2+0^2+0^2,\ 1=0^2+0^2+0^2+1^2,\ 2=0^2+0^2+1^2+1^2,\ \ldots$$
$$6=0^2+1^2+1^2+2^2,\ 7=1^2+1^2+1^2+2^2,\ 8=0^2+0^2+2^2+2^2,\ \ldots$$
$$23=1^2+2^2+3^2+3^2,\ 24=0^2+2^2+2^2+4^2,\ \ldots$$
$$71=2^2+3^2+3^2+7^2,$$

and so on.

It seems never to stop at all! In fact it never does stop. According to a famous theorem first proved in the eighteenth century by the great French mathematician Joseph L. Lagrange, *every* number is, indeed, the sum of four squares. It is not such an easy theorem. Even Lagrange's contemporary, the great Swiss mathematician Leonhard Euler, a man of astounding mathematical insight, originality and productiveness, had tried but failed to find a proof, so I am certainly not going to trouble the reader with the details of Lagrange's argument here.

Instead, let us try a calculation that is very much easier to see never to stop.

> *Find the smallest odd number that is the sum of* n *even numbers.*

The poor computer that is set mindlessly upon this task will certainly never complete its work, no matter what n is – because even numbers always add to even numbers.

I have given some examples of calculations, some of which will eventually terminate to produce an answer and some of which continue for ever. How are we to decide which of these two possibilities will occur in any particular case? When a calculation does not ever stop, by what means can we ascertain this fact? We have seen that this may be hard, as was the case with Lagrange's theorem, but sometimes it is easy, as in the last example. Are mathematicians themselves using some calculational procedure in order to ascertain that non-stopping calculations actually do not stop?

Let us imagine that they do use such a procedure, and, moreover, they are aware of the nature of this procedure and of the fact that the procedure that they use is sound – that is to say, that it does not erroneously come to the conclusion that a calculation does not stop when in fact it does. It will not be necessary to assume that this procedure can, in every case, ascertain that a non-stopping calculation does not in fact stop.

Let us call our putative procedure A. Then when A is presented with a calculation C and with the number n on which C acts, it will be set into action. If the calculation A itself successfully comes to a halt, then it will have decided that $C(n)$ does not in fact terminate. (Note that A is not a procedure for deciding that calculations *do* terminate. We might have some other procedure B for that kind of decision. If we want to incorporate B into A, we can do so by employing the device of putting A into a 'loop' whenever B successfully comes to its conclusion, thus making sure that the 'A' used in the argument will actually not terminate when the calculation does. This is just a technical point. I mention it only because sometimes people are disturbed if A ignores arguments that show that a calculation will stop.)

In order to be a little clearer about how a calculation (A) can act on another calculation (C), let us specify the various calculations C by giving each one a separate number. Thus the different calculations will be

$$C_0, C_1, C_2, C_3, C_4, C_5, \ldots$$

where we can think of this ordering as being provided by the numerical ordering of the computer programs that specify these calculations in turn. Technically, C_r could be the 'r^{th} Turing machine' in some standard system of numbering. We can now think of A as a calculation acting on the two numbers r and n, and conclude

If A*(r,n) stops, then* C$_r$*(n) does not stop*

The calculation A is still not quite of the form of the other calculations that we have been considering since it acts on two natural numbers, not one. Let us remedy this by considering only the cases for which $r=n$. (This perhaps seems an odd thing to do, but it is the crucial step in Gödel's and Turing's argument, itself taken from the famous 'diagonal slash' of the highly original nineteenth-century mathematician Georg Cantor.) We obtain

If A*(n,n) stops, then* C$_n$*(n) does not stop*

Now, A(n,n) is of the form of the calculations that we have been considering, so it must be one of them, say the k$^{\text{th}}$ one, and we have

$$A(n,n) = C_k(n)$$

and therefore, putting this in the displayed statement above

If C$_k$*(n) stops, then* C$_n$*(n) does not stop*

Taking the particular case $n=k$, we obtain

If C$_k$*(k) stops, then* C$_k$*(k) does not stop*

From this we deduce that C$_k(k)$ certainly will not stop (because if it did, then it doesn't!).

The remarkable thing, here, is that although we have ourselves just seen that C$_k(k)$ does not stop, the calculation A is incapable of ascertaining this fact. For A(k,k) is the same as C$_k(k)$, so if the latter does not stop, the former cannot stop either. Thus, A cannot successfully come to the conclusion that C$_k(k)$ does not stop! Since we have actually just established that C$_k(k)$ does not stop, it follows that the mathematical procedures that we use in order to establish that calculations do not stop are not accessible to the calculational procedure A. We note this would apply whatever A is, provided that we know what A is, and we know A to be sound. The inescapable conclusion seems to be:

> *Mathematicians are not using a knowably sound calculational procedure in order to ascertain mathematical truth*

We deduce that mathematical understanding – the means whereby mathematicians arrive at their conclusions with respect to mathematical truth – cannot be reduced to blind calculation!

DISCUSSION OF THE IMPLICATIONS OF THE GÖDEL ARGUMENT

I should address some of the possible loopholes and objections that various people have made to arguments of the type that I have just given.

A common reaction to the Gödel argument is simply not to take it seriously, for 'how could an argument of that kind possibly have anything to say about the mind?' But although this may be a natural reaction, it is no answer to the argument. If one believes that the conclusion is wrong, then one must find a flaw in the argument.

A worry that people often have is that I have given the argument in terms of a single A, whereas there might be a whole host of calculational procedures that mathematicians use. However, this is not a real objection. There is no difficulty about combining together many different such procedures (even an infinite number of them) into a single 'A', provided that it is a calculational matter to decide which procedure to use. It is only for convenience that I have phrased things as I have, and there is no loss of generality involved.

One common objection is to point out that the Gödel(–Turing) argument is itself something that one could envisage putting on a computer. There is nothing non-computable about generating all the steps of the argument as I have given them and, if we wish, we could include some version of the Gödel argument into our rules for deciding that calculations will not stop. But if we simply adjoin this new argument to the 'A' that we had before, we are really cheating, because that 'A' was already supposed to represent the totality of the means that are available to mathematicians for ascertaining that calculations do not stop. If we accept the Gödel argument as a new means for ascertaining that certain calculations do not stop, then our previous 'A' did not represent that totality. Instead, we should be using some new 'A', say A*, that includes this version of the Gödel argument. But if A* is supposed to represent that totality then we can apply our argument instead to A*, and again we obtain a contradiction. The point is that we cannot put the entire idea of the Gödel argument into calculational form even though we can incorporate certain instances of it into a calculation.

There is a closely related objection that people sometimes try to make against the version of the Gödel argument that I have given. The claim is sometimes made that the argument applies only to the one particular A that has been singled out, and that it is not a general objection against all As. This is a misconception about how the argument is being used, however. The argument has the form, familiar to mathematicians, of a *reductio ad absurdum*, whereby a hypothesis is put forward (here that there is some knowably sound calculational procedure that we use – and that we are calling it A) from which a contradiction is obtained, thereby showing that the hypothesis was false. The argument indeed rules out all such As, not just a particular one.

The Gödel argument is more often phrased in terms of some axiom system F, and in terms of the provability of mathematical results from F. Gödel's most familiar theorem shows that provided F is (and is believed to be) consistent – so it cannot be used to prove that a statement is true and false at the same time – then there are mathematical propositions that are (and are seen to be) true but which cannot be derived from F. The argument is often made that we cannot actually see that these propositions are true unless we can show that the axioms are consistent. I have not phrased my own argument in this way, but have referred to the soundness of the procedure A. If we trust A not to make mistakes, then we see, by the Gödel argument that A cannot represent the totality of our mathematical insights, whatever A might be. Likewise, if we trust F – and this implies that we believe F to be consistent, for otherwise we could use it to prove a nonsense like 1=2 – then we see that F does not represent the totality of our mathematical insights. If we do not trust F (or A) then it certainly cannot represent our insights!

Another objection I have seen made is that we cannot be sure that the numbers that we are talking about are actually the natural numbers 0, 1, 2, 3, 4, 5, . . . , but they might be some funny kind of 'surreal' numbers, where some of the things that would be true for the natural numbers turn out to be false for the surreal numbers. Although I cannot really take this argument seriously – because the numbers that we are talking about *are* the natural numbers, and not anything else – this objection contains, in a sense, the 'nub' of the mystery. How do we know that it is the natural numbers that we are indeed talking about? We cannot, merely by specifying a finite system of axioms or rules, completely distinguish the natural numbers from all the various kinds of 'surreal' numbers. Yet every child knows what the things 0, 1, 2, 3, 4, 5, . . . mean, despite this fact. Somehow we have a direct intuition which tells us what the 'natural number' concept is, given only very inadequate hints in terms of 'two bananas, five oranges, zero socks' and so on.

Let us then accept the apparently inescapable implication of the Gödel(–Turing) argument: *mathematicians do not simply ascertain mathematical truth by means of knowably sound calculational procedures*. There remain the possibilities that they might use unknowable or unsound calculational procedures – or, as is my own belief, that they simply do not just use calculational procedures when they ascertain truth. With regard to the calculational possibilities, I should point out that mathematicians certainly don't *think* that they are using unknowable or unsound procedures in order to ascertain mathematical truth! They are of the opinion that they are perfectly aware of what they are doing when they use whatever methods they use and, moreover, that these methods are perfectly sound. It is undoubtedly true that mathematicians make mistakes from time to time, but these mistakes are recognisable as such. Another mathematician might point out the mistake, or the very mathematician who made the mistake might notice it later. It is not that there are inbuilt errors that mathematicians are completely incapable of seeing as errors. Consciously, the methods that mathematicians use are neither unknowable nor unsound. If they are indeed using a horrendously complicated unknowable calculational procedure X, or an unsound calculational procedure Y, then these things would have to be completely unconscious.

Is it plausible that they are actually using such an X or Y without knowing it? One point should be emphasised here, and that is the apparently universal nature of the criteria that mathematicians use to establish the truth of their results. Suppose that each mathematician used a different X or Y, personal to that particular mathematician, then they would not be able to convince one another of their arguments. We require a universal X or Y that would have to be built into their brains in a way that would be common to all. How could such an X or Y have arisen? It would have to have been by means of the powerful processes of natural selection that Darwin himself revealed to us. But anyone who has glanced at any respectable modern mathematics research journal will realise how far-removed from the activities of the outside world are the deliberations of mathematicians. If it were the horrendously complicated unknowable X, or the complicatedly erroneous Y, that somehow got implanted in our brains, via our genes, it is very hard to see how this could have been by the process of natural selection, a process geared to promote the survival of our primitive remote ancestors. Much more likely is that there is no such X or Y but, instead, it is a non-calculational quality – the ability to *understand* – that natural selection has favoured. This quality is in no way specific to mathematics, but would have been immensely valuable to our ancestors in many different ways, providing a powerful selective advantage. Only incidentally does it turn out that this same quality is what is needed for mathematics.

WHAT UNDERLIES NON-COMPUTATIONAL BRAIN ACTION?

If we accept that we do something beyond computation when we understand, how can this be reconciled with the view that our brains are just physical objects governed by precise physical laws? One way out might be to adopt a mystical viewpoint according to which the behaviour of the mind could not be accounted for simply in terms of the physical brain. Apparently Gödel himself felt driven to this kind of solution.

For myself, I reject mysticism in favour of a scientific explanation. There are various possibilities to consider. For example, perhaps we use a calculational procedure that is continually improving itself. If so, then, we must ask, how is this improvement coming about. If the improvement is itself governed by some preassigned mechanism, then, as was the case with a neural network, it is still calculational. Might the improvements come about via some continual interaction with the environment? But if this is to give us something beyond calculation, it would imply that there is an essential feature of our environment that cannot even be *simulated* computationally. Of course, it may not be feasible to simulate the particular environment of a specific individual, but to suggest that it is in principle impossible to simulate any appropriate plausible environment is to suggest that there is something essential in the physical action of the world that lies beyond calculation. Once that possibility is accepted, then the possibility that our very brains might act according to some non-calculational action must also be allowed.

What about random ingredients? Would they count as 'non-calculational'? In the sense that a strict Turing machine does not allow for such ingredients, their inclusion would, indeed, take us out of calculational activity. However, in practice, purely random ingredients would add nothing useful to pure calculation. In fact, there are many calculational procedures that call for the inclusion of random ingredients, but these are usually implemented in practice by incorporating what are called 'pseudo-random numbers', these being numbers that are generated by some suitably complicated process that gives them the appearance of randomness even though they are not strictly random. For practical purposes randomness gives us nothing useful that cannot be achieved purely calculationally. (Closely related is the behaviour of what are called 'chaotic systems' which have the appearance of randomness even though they are entirely calculational.)

Finally, there is the possibility that, in appropriate circumstances, the actual behaviour of physical systems might be essentially non-calculational, a conscious brain being one example. My personal belief is that this is indeed the case, but there are several speculative elements that are involved in such a belief. First, one must ask where in physics non-calculational action might be found. I believe that such action must be in an area where present-day physics is in need of radical improvement – what is referred to as 'the measurement problem' in quantum theory. Roughly speaking, such an improved theory would supply a more satisfactory link between the micro-level of atoms and molecules (the 'quantum level') and the macro-level of discernable phenomena (the 'classical level'). I believe that brain action will never be properly understood without such a theory. At least something of this nature will be needed in order to explain the non-calculational aspects of mathematical and other kinds of understanding.

FURTHER READING

Boden, Margaret A., *The Creative Mind: Myths and Mechanisms*, London: Weidenfeld 1990.

Broadbent, D. (ed.), *The Simulation of Human Intelligence*, Oxford: Blackwell 1993.

Chou, Shang-Ching, *Mechanical Geometry Theorem Proving*, Dordrecht: Reidel 1987.

Nagel, E., and Newman, J.R., *Gödel's Proof*, London: Routledge and Kegan Paul 1958.

Penrose, Roger, *The Emperor's New Mind*, Oxford: Oxford University Press 1989. Vintage paperback, 1990.

232

IS CONSCIOUS AWARENESS CONSISTENT WITH SPACE-TIME DESCRIPTIONS?

Philosophy, mathematics and modern physics: A dialogue. Eds E. Rudolph and I. O. Stamatescu, 34–47. Springer, Berlin, 1994

Is conscious awareness consistent with space-time descriptions? *Philosophy, mathematics and modern physics: A dialogue.* Eds E. Rudolph and I. O. Stamatescu, 34–47. 1994

Is Conscious Awareness Consistent with Space-Time Descriptions?

Roger Penrose

Mathematical Institute, University of Oxford, Oxford, UK

1. Consciousness and Computation

It is a fact of nature that certain objects inhabiting this physical world are capable of evoking conscious awareness. Among such objects are healthy living human brains (at least when attached to human bodies). We do not know what other physical systems might be capable of this feat. There is a point of view which asserts that a sufficiently powerful computer, suitably programmed, could also evoke awareness. Indeed, it is not infrequently maintained that this conclusion would be a necessary implication of any completely scientific viewpoint. It is argued that the brain's function is simply "information processing" – the activity that computers also indulge in, sometimes very much more effectively than human brains. Although there are many activities in which the brain remains a more effective instrument than any computer system constructed to date, it is only a matter of time – so the argument runs – before computers will vastly exceed the capabilities of brains in all significant respects. Of course, opinions might differ widely as to the timescales involved. Some say that we shall be superseded in less than forty years (cf. Moravec 1988), whilst others argue that it will be many centuries, perhaps even millenia, before computers will overtake us, though eventually they must do so – if we do not destroy ourselves first!

Others (cf. Searle 1980, 1987) argue that the mere computational activity that computers indulge in would not, of itself, evoke awareness or any other mental quality. In particular, the quality of "understanding" would not be present by virtue of the mere carrying out of some calculation. Nevertheless, the possibility of a computational *simulation* of the activity of a brain while in the process of doing whatever it does when a person actually understands something is not denied by such people. But such a simulation is argued as being not *itself* capable of evoking any actual understanding or any other conscious mental quality.

The first of these viewpoints is varioulsy referred to as "functionalism" or as "strong AI" (where AI stands for "artificial intelligence"). This basic position is encapsulated in the assertion:

$\mathcal{A}$. All thinking is computation, and the mere carrying out of appropriate computations will evoke feelings of conscious awareness.

The second viewpoint can be summarized as:

$\mathcal{B}$. It is the brain's physical action that evokes awareness; and any physical action can be simulated computationally; but computational simulation by itself cannot evoke awareness.

It should be pointed out, however, that there is really no distinction between viewpoints $\mathcal{A}$ and $\mathcal{B}$ with regard to the question of the eventual superiority of computers over human brains. If an effective total simulation of human intelligence can be achieved – and then vastly exceeded – by some robot, then these achievements would be in no way hindered if the robot did not "actually" possess any awareness. Paradoxically, it is $\mathcal{A}$ rather than $\mathcal{B}$ that provides us with ultimate hope for a future for human awareness. If the mentality that each of us possesses is something that can be translated into some computer program, as is asserted by viewpoint $\mathcal{A}$, then one can envisage this program being transferred to the supercomputer that controls a robot, so that the person's actual awareness would itself be transferred to robot form (Moravec 1988). Viewpoint $\mathcal{B}$ does not admit this possibility, and it leaves us with the prospect of our planet's ultimate domination by totally insentient robots.

There is, however, an important assumption involved in $\mathcal{B}$ (and also in $\mathcal{A}$). This is that physical actions can indeed always be simulated computationally. I shall argue, in a moment, that the quality of conscious "understanding" is something that cannot even be *simulated* computationally. This goes beyond Searle's argument that no computational simulation could, of itself, evoke the internal mental quality of "understanding" since I claim that *outward* manifestations of this quality would differ from the outward manifestations of any computer-controlled robot. Thus, according to my own point of view, there must be something in the physical action of a brain, when it is in the process of consciously understanding something, that defies any computational description whatever. Accordingly, my own viewpoint asserts:

$\mathcal{C}$. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.

Some might regard such a possibility as being totally outside the bounds of science. However, there are many things within precise mathematical descriptions that to defy computation. For example, there is no computational procedure for deciding whenever it is the case that a family of polynomial equations, with integer coefficients, in several integer variables, has no solution. Likewise there is no computation that can decide whenever a given set of polygonal tile shapes (say made up of squares joined together, for simplicity) is capable of tiling the entire Euclidean plane. Nor is there any computational procedure that correctly decides whenever two four-dimensional manifolds are topologically distinct from one another. It is certainly possible to produce "toy model universes" based upon such algorithmically insoluble mathematical structures (cf. Penrose 1989, p. 170). Non-computable physical laws, though unfamiliar to us, are certainly within scientific possibility. Moreover, in accordance with viewpoint C, they might underlie whatever physical actions are taking place when a brain indulges in conscious thinking. Thus, we do not need to be driven outside the bounds of science when rejecting viewpoints $\mathcal{A}$ or $\mathcal{B}$.

In a moment I shall give what I believe are the most scientifically compelling reasons for indeed rejecting these two viewpoints. There does not appear to be any action within known physical laws that definitely requires a non-computational description, however. My own standpoint, as I shall elaborate a little at the end of this article, is that this action has to do with whatever is physically going on in the process that we unsatisfactorily describe as "the collapse of the wave function". It is my opinion that a fundamental advance in physical understanding is needed before this process can be properly described in scientific terms – and that non-computational ingredients will have to be introduced as an integral part of this advance. I must emphasize that in suggesting this I am *not* being driven to taking a non-scientific attitude, as some would maintain, such as is embodied in the alternative final viewpoint:

$\mathcal{D}$. Awareness cannot be explained in physical, computational, or any other scientific terms.

2. Non-computational Nature of Mathematical Understanding

In order to show that the conscious mind is capable of *some* kind of non-computational activity, I shall have to turn to mathematics. Once it is established that conscious *mathematical* understanding is something that cannot be simulated computationally, then my main point will have been established. It is not likely that there is anything special about mathematics in this respect. If conscious mathematical activity is seen to be something that cannot be simulated computationally, then this opens up the possibility that perhaps *all* conscious brain activity depends upon a physical action that is non-computational.

Let us consider computations in general. What is computation? It is essentially anything that can be carried out by a modern general-purpose computer. Technically this is what we call an "algorithm" or the action of a "Turing machine". Strictly speaking, such action might require an unlimited storage capacity, but I am not going to worry about this proviso here. Sometimes people consider that certain types of computational activity might come outside what is normally called "algorithmic", and they cite such things as "parallel action", "connectionism" (neural networks) "learning systems", "heuristics", "random elements" and input from the "environment". In my discussion, I am considering all these things as being effectively *included* in what I mean by computation, since they can (and often are) simulated on an ordinary general purpose computer. This certainly applies to parallel action, even for those procedures for which a parallel architecture is much more effective than a serial one. Well-defined calculational rules can also be given to simulate "learning" – such as that which underlies the action of neural networks – and also "heuristics". Random elements can be effectively simulated by using pseudo-random numbers, these being generated by perfectly clear-cut algorithms. Moreover, the environment can, in principle, also be effectively simulated – unless the relevant physics governing the behaviour of that environment is itself in some way non-computable, which is what the argument is eventually trying to establish in any case.

An important fact about computations is that some of them do not ever terminate. Consider the example:

> "Find an odd number that is the sum of two even numbers".

One could clearly set a computer to perform this task. For example, it could consider all the triples (a,b,c) one after the other, ordered appropriately, where a, b and c are natural numbers (i.e. non-negative integers) for which a is odd, and b and c are even, stopping only when a triple is found for which a = b+c. This task would obviously never end. A considerably less obvious example of a non-terminating computation is:

> "Find a number that is not the sum of four square numbers".

(By a "number", I here mean, as above, a "natural number": one of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, \ldots$.) The fact that this computation does not stop is far from obvious. It follows from a theorem in number theory, proved by Lagrange in the eighteenth century. If, instead, the computation had been:

> "Find a number that is not the sum of three square numbers";

then the computation would have stopped at 7 (which needs four squares: $7 = 1^2 + 1^2 + 1^2 + 2^2$). For another example, we could consider:

> "Find an even number greater than 2 that is not the sum of two prime numbers".

It is not known whether this computation will ever stop. The assertion that it would not is the famous Goldbach conjecture, unsettled to this day.

How do mathematicians convince themselves that certain computations in fact do not stop? In the first of our examples, we could say that this fact is "obvious", and if desired we could provide an appropriate proof; whilst in the second, the reasoning is quite involved – though it can be followed by any mathematician who follows through the details of Lagrange's proof.

When they come to their conclusions in such matters, are mathematicians always simply following some algorithm – or computation – acted out by the neurons inside their heads? I want to present an argument that renders this possibility exceedingly unlikely. It is based on the ideas of Gödel and Turing, put forward in the 1930s – basically it is a version of Gödel's famous incompleteness theorem – but only an extremely simplified distillation of these ideas will be needed here.

First, for technical reasons, we shall need to consider computations that can apply uniformly to an arbitrary natural number n, rather than having just single statements, for example:

"Find an odd number that is the sum of n even numbers";

or

"Find a number that is not the sum of n square numbers".

The first of these stops for no value of n whatever, whilst the second stops only for $n = 0, 1, 2, 3$. Now such computations can certainly be listed (computably) in some order:

$$C_0(n), C_1(n), C_2(n), C_3(n), C_4(n), C_5(n), \ldots .$$

(For example, we can take computer programs written in some standard computer language, listing them in order of length, and "alphabetically" within each specified length.) Suppose that we have some computational procedure P that, upon coming to an end, provides us with a demonstration that certain of the above computations actually do *not* terminate. The procedure P will take note of both the particular calculation C_a that it is examining and the number n that this calculation is applied to, so it depends upon two natural numbers a and n: we write this $P(a, n)$. Thus we have:

if $P(a, n)$ terminates, then $C_a(n)$ does not terminate.

It is not required that the computational procedure P be capable of *always* deciding that a non-terminating action $C_a(n)$ does not in fact terminate, but we do want to have been convinced that the procedure P does not ever give us wrong answers, i.e., that $P(a, n)$ never comes to an end when $C_a(n)$ actually does terminate. Thus, we are to think of $P(a, n)$ as some known procedure that is *sound* (and that we believe to be sound) which, when it successfully terminates, provides us with a proof that $C_a(n)$ does not terminate. We shall try to imagine that P is in fact a formalization of the totality of all the procedures available to human mathematicians for deciding that computations do not terminate, but the argument has significance *whatever* sound procedure P might happen to be. Next, consider the computation $P(n, n)$, where I have put a equal to n. This is a computation depending upon just the one variable n, so it must be in our original list; say it is the k^{th} one:

$$P(n, n) = C_k(n).$$

Now take the particular value of n where $n = k$, so we have, from the above (with $a = n = k$):

if $P(k,k)$ terminates, then $C_k(k)$ does not terminate.

We shall try to see whether $P(k,k)$ actually terminates or not. Suppose that it does; then by the above, we see that $C_k(k)$ does not terminate. But from the displayed line before (with $n = k$), $C_k(k)$ is in fact identical with $P(k,k)$, so $P(k,k)$ does not terminate after all. This must be the right answer: $P(k,k)$ does *not* terminate, and so $C_k(k)$ does not terminate either.

The remarkable conclusion that we have drawn is that we have been able to find a calculation (namely $C_k(k)$) that *we* can see does not terminate (by the above argument), yet the calculational procedure P is unable to ascertain this fact (since $P(k,k)$ does not terminate)! Thus if P had actually been a formalization of all procedures available to human mathematicians for deciding that computations do not terminate – as we were invited to imagine above – then we should seemingly have arrived at a contradiction. We are able to see that $C_k(k)$ does not terminate whereas P fails to ascertain this fact. This conclusion depends upon our actually *knowing* the computational procedure P and, moreover, knowing that P is actually *sound*. We deduce that human mathematicians are not using a knowably sound computational procedure to ascertain mathematical truth! The mathematical insights that are in principle available to us are not things that can be reduced to computation.

There are several objections that can be (and have been) raised against this kind of argument (e.g., the many criticisms of a related discussion given by Lucas 1961; cf. Good, 1967, 1969, Benacerraf 1967, Bowie 1982, Hofstadter 1979, Lewis 1969, 1989). Some of these are misconceptions that do not really apply to the particular form of *reductio ad absurdum* that I have presented here. Others represent possible viewpoints that try to come to terms with the conclusion as I have actually presented it. Let me very briefly address these viewpoints in turn. (A much more thorough analysis will be given in a forthcoming book that I am at present engaged in writing.)[1] We might perhaps use (i) a horrendously complicated unknowable computation X, or (ii) an unsound but almost sound computation Y (which was probably Turing's own preferred solution), or (iii) an ever-changing computation, or (iv) random elements, or (v) input from the environment. On the other hand, perhaps we are indeed driven either to believing in: (vi) some mystical concept of "mind" lying outside of physical explanation (which may have been Gödel's solution) – viewpoint $\mathcal{D}$ – or else, by insisting that we remain within the realms of scientific explanations, we find ourselves driven to my own conclusion: (vii) that our (conscious) thought processes depend upon some non-algorithmic physics, a physical action that cannot even be effectively *simulated* by any computation – viewpoint $\mathcal{C}$.

---

[1] *Shadows of the Mind*, Oxford University Press, to be published

My objections to (iii), (iv) and (v) above have been outlined earlier. My difficulties in accepting (i) or (ii) stem partly from the fact that the way that we actually ascertain mathematical truth seems to be totally unlike the use of a horrendously complicated unconscious computational procedure X or of a fundamentally incorrect computational procedure Y. We must always be able in principle to break down our mathematical arguments into elementary steps that are "obvious" to our conscious understanding of the concepts involved. The discussion that I have given above shows that we cannot ever accurately and knowingly reduce such understanding to a computation. Once we have perceived how to encapsulate a body of sound mathematical insights into a computational procedure, then this very perception provides us with a *new* mathematical insight, as trustworthy as those insights that we have incorporated into our computational procedure, but which is not itself incorporated.

Let me put things another way. The most familiar form of the Gödel argument provides a statement that asserts the consistency of a set of axioms, this statement being not derivable from the axioms themselves provided that those axioms are actually consistent. If we believe that the axioms are *sound* (i.e., that the deductions from them are *true* mathematical statements), then we must believe that they are indeed consistent, so we must believe in the truth of the Gödel proposition. This insight, that enables us to deduce from soundness of a system of axioms that they must in fact be consistent, is an insight that cannot be incorporated into the axioms themselves. In general terms, it is our *understanding* of what we are doing and of what the terms of our mathematical expressions actually *mean* that enables us to break out of any formalized computational procedure. We *see* the truth of the Gödel proposition by use of our understanding of the meanings of the terms involved. People often object to using the Gödel procedure as a demonstration that mathematical insight goes beyond any computational scheme, saying that we cannot "see" the truth of the Gödel proposition unless we already know that the axioms are consistent. But the point is not that we are supposed to be able to "see" the truth of the Gödel proposition for an *arbitrary* system of axioms; it is only that we can do this for a system of axioms that we are already accepting as sound – and therefore consistent – because, by the hypothesis of the *reductio ad absurdum*, this axiom system (like the computational procedure P) is being hypothesized as representing the totality of the acceptable insights that are available to human mathematicians. The argument shows, quite rigorously, that one cannot form any axiom system that represents, and is believed to represent, the totality of sound mathematical insights that are available to human mathematicians.

The argument does allow, however, that we might be using an unconscious axiom system – equivalently, an unconscious algorithm – in order to derive our mathematical results. This system (algorithm) would have to be either unsound (Y above) or, if sound, its soundness would have to be unknowable to us (X above). Both possibilities strike me as being most implausible, for the process actually underlying our mathematical beliefs. In connection with Y, we should note that although mathematicians do make mistakes, these can be seen to be mistakes by the very people who make them. Once a mistake has been pointed out to a mathematician then, at least in principle, that mistake can be recognized as such. This is quite different from the blind following of an unsound algorithm. With regard to X (or Y), we should bear in mind that most of the mathematics that modern mathematicians are concerned with is enormously far from everyday experience. It is very hard to see how either X or Y could have arisen by natural selection, when our ancestors were presumably concerned with much more mundane matters (such as how to catch mammoths, etc.). How could natural selection have produced a highly sophisticated unconscious unknowable algorithm, an algorithm common to everyone in the mathematical community, that enables mathematicians to decide the truth of obscure statements which have, for the most part, almost no relevance whatever to survival? It is quite possible, however, that a general faculty for (conscious) *understanding*, be it something that could be employed towards the survival of mankind (e.g., the construction of mammoth traps) or something that ultimately was found could be used for doing mathematics (at *all* levels of sophistication), has been produced by natural selection. So long as it is accepted that the quality of understanding can be something *non-algorithmic*, then the Gödel argument is satisfied.

3. Where in Physics May We Find Non-computational Action?

I believe that these arguments indeed lend strong support to viewpoint $\mathcal{C}$ – assuming that we are attempting to follow the scientific path towards an understanding of the phenomenon of consciousness, and thereby rejecting $\mathcal{D}$. What does this say about the physical laws that control the action of a conscious brain? It seems to me that the very least that we can conclude is that these laws must be fundamentally non-computational. By this, I do not simply mean that their consequences are unpredictable in the sense of "chaos", according to which one would need to know initial conditions to an absurdly unreasonable precision in order to predict the future behaviour of the system. As far as one can tell, chaotic systems merely introduce an effectively random element into the future behaviour; and as we have seen, randomness is not useful for taking us outside the bounds of what can be achieved by a Turing machine. After all, when one uses pseudo-random number generators, one is effectively calling upon this same feature of chaotic systems. The numbers appear to be random merely because they are "unpredictable" in very much the same sense that chaotic systems are "unpredictable".

There does remain a possibility that when the behaviour of chaotic systems is examined more thoroughly, it will be found that rather than just appearing random, they will begin to exhibit some genuinely useful non-computable behaviour, but I have seen no indications of this as yet. Much more likely, to my way of thinking, is that usefully non-computational behaviour will emerge as a necessary ingredient of an improved quantum theory, in which the unsatisfactory procedure of "wavefunction collapse" or "state-vector reduction" – which I shall denote by the letter **R** – is replaced by a mathematically sensible dynamical process. In my opinion, there is a fundamental gap in our present physical understanding between the small-scale "quantum level" (which includes the behaviour of molecules, atoms and subatomic particles) and the larger "classical level" (of macrosopic objects, such as cricket balls or baseballs). Applying at the quantum level are the procedures of unitary evolution – which I denote by the letter **U** – as can be described very precisely by Schrödinger's differential equation. But when we pass to the classical level, it seems that the world does *not* evolve according to Schrödinger's equation – a *linear* equation which preserves the superposition of "alternative worlds" that is a feature of descriptions at the quantum level. Instead, we adopt the probabilistic procedure **R**, which selects, in some appropriately random way, one alternative from amongst the multitude that are taken as co-existing in quantum-level superposition.

The issue of whether a gap in our physical understanding exists in relation to this issue is a matter of *physics* – unrelated, at least in the first instance, to the question of "minds". I believe that there is good reason, on purely physical grounds, to believe that some fundamentally new understanding is indeed needed here. But amongst physicists as a whole, this entire issue is a matter of dispute at the moment, and it would have to be admitted that the majority view would appear that *no* new theory is needed (although such outstanding figures as Einstein, Schrödinger, deBroglie, Dirac, Bohm and Bell have all expressed the need for a new theory). It is my own strong belief that a radical new theory is indeed needed, and I am suggesting that an essential component of this theory will be its fundamentally non-computational character (cf. Penrose 1989).

An important feature of any "realistic" modification of quantum theory, in which the procedure **R** is replaced by some well-defined physical action, is that this action must be *non-local* in a way that is consistent with the type of violation of Bell's inequality that has been observed in actual experiments (cf. Aspect 1976, Aspect and Grangier 1983). In these particular experiments, pairs of photons are emitted in opposite directions from a source (where the source consists of atoms that undergo transitions from one energy level to another), the photons being individually detected in two separate places, some twelve meters apart, at opposite sides of the source. The detectors are arranged to measure the components of the photons' polarization at certain orientations, where the precise choice of orientation is not made until the photons are in full flight from source to detectors. It is found (in accordance with Bell's 1964 analysis of such situations) that the statistical correlations between the polarization measurements at the two detectors is inconsistent with the pairs of photons being separate independent entities, but each pair of photons consists of a single "connected" entity, at least up until the choice polarization measurement is made on one of the photons or the other. This is despite the fact that the two photons are many meters apart when this choice is indeed made.

The standard quantum-mechanical "picture" is of a state vector that describes each photon *pair* as evolving according to U – the deterministic computable Schrödinger equation – up until the moment that a measurement is first made on a member of the pair. At that moment one evokes the *second* quantum-mechanical procedure R (statevector reduction), which is no longer deterministic, but *probabilistic*, according to standard theory. A measurement of one photon's polarization will immediately force the polarization of the *other* photon into a particular state, which is the one that is then measured by *that* photon's detector. Somehow, it would seem that a "message" is sent instantaneously from one photon to the other, signalling how that other photon is to arrange its polarization. This "message" does not transmit actual information concerning the choice of polarization that is made in the first measurement – in the sense that one cannot use this effect to send an instantaneous signal from one place to another – but without postulating some kind of "message" of this kind, one cannot make sense of the statistics that relate the polarization measurements on the two photons. Instead of thinking in terms of "messages", one can think of the two photons as being, in a sense, one "connected entity", but this really amounts to a restatement of the same thing.

An essential difficulty with such non-locality arises if we try to make a relativistically invariant picture of what is happening. With superluminary "messages" of this kind, it is not clear which photon is sending the "message" and which is receiving it. In the standard quantum-mechanical state-vector description, it is the first photon measured that determines the *time* at which the R procedure is enacted, so the other photon encounters this "reduction" phenomenon *before* it is actually measured. But in another Lorentz frame, things can be the other way about, and it is now the photon that had previously been called the "first" photon that encounters R before it is actually measured. It should be made clear, that the actual probabilities come out the same whichever Lorentz frame is used, so there is no physical *contradiction* with relativity here. Nevertheless, the situation poses a profound challenge to any "realistic" space-time picture of what is going on, consistently with the tenets of relativity theory. Relativity tells us that physical effects are supposed not to propagate faster than light signals. The mysterious quantum *entanglement* that exists between the two photons is something that indeed defies a normal relativistically invariant space-time description – though various "abnormal" descriptions are possible. The simplest such "abnormal" description that I have come across involves *two* state vectors to describe any system, one of which evolves forward in time, from a measurement, in the normal way whilst the other evolves *backwards* in time. (This was described to me by Yakir

Aharonov in 1991). Other possibilities that I have encountered seem no less strange.

Whatever the merits of such proposals, it is my own strong opinion that when a sensibly "realistic" picture of what is involved in the quantum-mechanical R procedure finally comes to light, it will be found that not only does this involve non-computable action of some kind, but that the description will indeed turn out to be at variance with our presently conventional space-time view.

4. Time and Conscious Perceptions

Returning to the issue of our conscious perceptions, it should be recognized that there is something very odd about our perceived "flow of time", from the point of view of modern physical theory. According to modern physics, "time" is merely a particular coordinate measure in the description of the location of a space-time event. There is nothing in the physicist's space-time descriptions that singles out "time" as something that "flows". Indeed, physicists quite often consider model space-times in which there is only *one* space dimension in addition to the single time dimension; and in such two-dimensional space-times there is nothing to say which is space and which is time. Yet, no-one would consider space to "flow"! It is true that one often considers time-evolutions in physical problems, were one may be concerned with computing the future from the present state of the system. But this is not at all a necessary procedure, and calculations are normally carried out this way *because* one is concerned with modelling, mathematically, our experiences of the world in terms of the "flowing" time that we seem to experience. It is our apparent experiences that tempt us to bias our computational models of the world in terms of time-evolutions – frequently, but not invariably – whilst the physical laws themselves do not contain such a compelling inbuilt bias.

Indeed, one could say that is only the phenomenon of *consciousness* that requires us to think in terms of a "flowing" time at all. According to relativity, one has just a "static" four-dimensional space-time (admittedly with some differential equations relating behaviour at different space-time events), with no "flowing" about it. So consciousness appears to have some special role in relation to the passage of time.

On the other hand, it would be unwise to make too strong an identification between the phenomenon of conscious awareness, with its seeming "flowing" of time, and the physicist's use of a real-number parameter t to denote what would be referred to as a "time coordinate". In the first place, it is clear that the precise concept of a "real number" is not completely relevant to our conscious perception of the passage of time, if only for the reason that we have no sensibility of very tiny time-scales – say time-scales of even just one hundredth of a second, for example – whereas the physicist's time-scales hold good down to some 10^{-25} seconds, or perhaps down to 10^{-53} seconds, and the mathematician's concept of time as a *real number* would require that there be *no* limit of smallness whatsoever, below which the concept meaningfully applies.

Quite apart from these issues, there are some puzzling experimental findings about the *timing* of conscious experiences that have almost paradoxical implications. In the mid-1970s, H.H. Kornhuber and his associates (cf. Deeke *et al.* 1976) used electroencephalograms (EEGs) to record electrical signals at a number of points on the heads of several human volunteers, who were asked to flex the index finger of their right hands suddenly at various times *entirely of their own choosing.* The idea was that the EEG recordings would indicate something of the mental activity that is taking place within the skull, and which is involved in the actual conscious decision to flex the finger. In order to obtain a significant signal from the EEG traces, it was necessary to average the traces from several different runs, and the resulting signal was not very specific. However, what was found was that there was a gradual buildup of recorded electric potential for a *full second*, or perhaps even a second and one-half, *before* the finger was actually flexed. This seemed to indicate that the conscious decision process takes a second or more in order to act. This is to be contrasted with the much shorter time that it takes to respond to an external signal if the mode of response has been laid down beforehand. For example, instead of it being "freely willed", the finger flexing might be in response to the flash of a light signal. In that case a reaction time of about one-fifth of a second would be normal – about five times faster than the "willed" action that is relevant to Kornhuber's data. Thus, according to these experiments, and also more recent experiments of a similar nature performed by Libet and others, it would appear that conscious "willed action" takes about a second in order to act.

In a second experiment, Libet (1979), in collaboration with Feinstein, tested subjects who had had to have brain surgery (for some reason unconnected with the experiment) and had consented to having electrodes placed in parts of the brain concerned with receiving sensory signals. The upshot of Libet's experiment was that when a stimulus was applied to the skin of these patients, it took about half a second before they were consciously aware of that stimulus, despite the fact that the brain itself would have received the signal of that the stimulus in only about a hundredth of a second, and a pre-programmed "reflex" response to such a stimulus (cf. above) could be achieved by the brain in about a tenth of a second. Moreover, despite the delay of half a second before the stimulus reaches awareness, there would be the subjective impression by the patients themselves that they had already been aware of the stimulus at an earlier time – essentially simultaneously with the moment that it had been applied.

Each of these experiments, by itself, would not be paradoxical, although maybe a little disturbing. Perhaps one's conscious decisions to act *do* need to occur something like a whole second before that action finally takes place; perhaps one's passive consciousness *does* require something like half a second before it can be evoked. But if we take these two findings together (and take each of them at its face value) then we seem to be driven to the conclusion that in any action in which an external stimulus leads to a consciously controlled response, something like a second and one half, at least, is needed before that response can occur.

This assumes that there is an "actual time" at which a conscious experience takes place, and that time must precede the "actual time" at which the conscious response to that experience occurs. This indeed would have seemed to have been our natural expectation. However, the Kornhuber-Libet experiments appear to lead us into some difficulties with this "natural expectation". For there are many instances of human activity that have the appearance of being consciously controlled, yet it takes far less than a second and one-half between the initial stimulus and the resulting response. Perhaps the most commonplace of these is in ordinary conversation, where consciously controlled response seems to take place in a fairly small fraction of a second.

No doubt there are many plausible loopholes in this apparent paradox, and perhaps nothing mysterious about the timing of consciousness need be inferred. My own suspicion, however, is that there could indeed be something unexpected and important underlying these experimental findings. Consciousness is already seen to have an odd and unique relationship with the (apparent?) passage of time. I believe that there may well be something else that is odd about this relationship, and it may not even be meaningful really to attach a "time" to a conscious experience. If it is the case, as I have being trying to argue above, that the phenomenon of consciousness is dependent upon some physical process that underlies the R procedure of quantum mechanics, and bearing in mind that the non-locality and non-causal aspects of this procedure are inherent in quantum-entangled systems, we should indeed be led to expect something odd in the relationship between consciousness and time.

I had better leave my speculations at that, for the moment. Perhaps further developments in theory and in physical and neurophysiological experiments will shed more light on these questions in the future.

References

Aspect, A. (1976): Proposed experiment to test the nonseparability of quantum mechanics, *Phys. Rev.* **D14**, 1944-51.

Aspect, A. and Grangier, P. (1986): Experiments on Einstein-Podolsky-Rosen-type correlations with pairs of visible photons, in *Quantum Concepts in Space and Time*, eds. R. Penrose and C.J. Isham (Oxford University Press, Oxford).

Bell, J.S. (1964): On the Einstein-Podolsky-Rosen paradox, *Physics*, **1**, 195–200.

Benacerraf, P. (1967: God, the Devil and Gödel, *The Monist* **51**, 9–32.

Bowie, G.L. (1982): Lucas' number is finally up, *J. of Philosophical Logic* **11**, 279–285.

Deecke, L. Grötzinger, B., and Kornhuber, H.H (1976): Voluntary finger movement in man, cerebral potentials and theory, Biolog. Cybernetics **23**, 99.

Good, I.J. (1967): Human and machine logic, *Brit. J. Philos. Sci.* **18**, 144–147.

Good, I.J. (1969): Gödels's theorem is a red herring, *Brit. J. Philos. Sci.* **18**, 359–373.

Hofstadter, D.R. (1979): *Gödel, Escher, Bach: an Eternal Golden Braid* (Harvester Press, Stanford Terrace, Hassocks, Sussex).

Lewis, D. (1969): Lucas against mechanism, *Philosophy*, **44**, 231–3.

Lewis, D. (1989): Lucas against mechanism II, *Can. J. Philos.*, **9**, 373–6.

Libet, B. (1979): Subjective referral of the timing for a conscious sensory experience, *Brain* **102**, 193.

Lucas, J.R. (1961): Minds, Machines and Gödel, *Philosophy* **36**, 120–124; reprinted in Alan Ross Anderson (1964): *Minds and Machines* (Prentice Hall, Englewood Cliffs).

Moravec, H. (1988): *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, Cambridge, Mass.; London).

Penrose, R. (1989): *The Emperor's New Mind* (Oxford University Press, New York, Oxford).

Searle, J.R. (1980): Minds, Brains and Programs, in *The Behavioral and Brain Sciences, Vol. 3* (Cambridge University Press, Cambridge).

Searle, J.R. (1980): Minds and Brains Without Programs, in *Mindwaves*, eds., C. Blakemore and S. Greenfield (Basil Blackwell, Oxford).

233

A TWISTOR-TOPOLOGICAL APPROACH TO THE EINSTEIN EQUATIONS

Twistor Newsletter **38** (1994)

This follows up further on Chapters 200, 202, 203, 205, and 230, but unresolved issues still remain.

A Twistor-Topological Approach to the Einstein Equations

In my articles in TN 31 (pp. 6–8) and TN 32 (1–5) (cf. also RP(1992)) the suggestion was made that the appropriate twistor concept for vacuum space-times is as a *charge* for *helicity* $\frac{3}{2}$ *fields*. This idea has been further persued by RP, TN 33 (1–6), RP(1994), LJM & RP, TN 37 (1–6), and JF, TN 37 (7–9). In particular, in RP(1992) and TN 32, I described how, in flat space-time $\mathbb{M}$, one may use a "chopping and pasting" method to work out the "twistor charge", where this charge appears as a "second kind" of gauge freedom for the potentials for the helicity $\frac{3}{2}$ field, namely that freedom in the gauge quantities which does not affect the potentials. This is best exhibited in terms of an exact sequence

$$0 \to \left\{\begin{matrix}\Omega\\ \Pi\end{matrix}\right\} \to \left\{\begin{matrix}\omega\\ \pi\end{matrix}\right\} \to \left\{\begin{matrix}\rho\\ \sigma\end{matrix}\right\} \to \left\{\begin{matrix}0\\ \psi\end{matrix}\right\} \to 0 \qquad \textcircled{A}$$

charges gauges potentials fields

(these actually being *local twistor* sheaves: LJM & RP, TN 37). I am here using the "Dirac gauge" for the potentials $\rho_{A'BC}$, $\sigma_{A'B'C}$, which implies that they are *symmetric*, as is the field $\psi_{A'B'C'}$:

$$\rho_{A'BC} = \rho_{A'(BC)}, \quad \sigma_{A'B'C} = \sigma_{(A'B')C}, \quad \psi_{A'B'C'} = \psi_{(A'B'C')},$$

and we have the equations

$$\nabla^{A'}_{A}\rho_{A'BC} = 0, \qquad \textcircled{B}$$

$$\nabla^{B}_{B'}\rho_{A'BC} = 2i\,\sigma_{A'B'C}, \quad \nabla^{B'}_{B}\sigma_{A'B'C} = 0, \qquad \textcircled{C}$$

$$\nabla^{C}_{C'}\sigma_{A'B'C'} = \psi_{A'B'C'}, \quad \nabla^{A'}_{A}\psi_{A'B'C'} = 0. \qquad \textcircled{D}$$

(The factor $2i$ is chosen for later convenience.) The gauge freedom is given by

$$\rho_{A'BC} \mapsto \rho_{A'BC} - i\varepsilon_{BC}\pi_{A'} + \nabla_{CA'}\omega^{B}, \quad \sigma_{A'B'C} \mapsto \sigma_{A'B'C} + \nabla_{CB'}\pi_{A'} \qquad \textcircled{E}$$

where

$$\nabla^{B}_{A'}\omega_{B} = 2i\,\pi_{A'}, \quad \nabla^{A'}_{B}\pi_{A'} = 0, \qquad \textcircled{F}$$

so $\pi_{A'}$ satisfies the Dirac-Weyl neutrino equation for helicity $1/2$, and ω^A is a "potential" for π (or, equivalently, a "neutrino" field of helicity $-\frac{1}{2}$ which has π as a source).

The "second kind" of gauge freedom, leaving ρ and σ alone, is given by

$$\omega_A \longmapsto \omega_A + \Omega_A, \quad \pi_{A'} \longmapsto \pi_{A'} + \Pi_{A'} \qquad Ⓖ$$

where

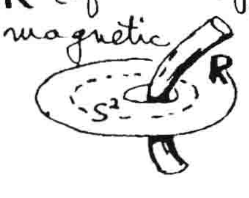
$$\nabla_{AA'}\Omega_B = -i\varepsilon_{AB}\Pi_{A'}, \quad \nabla_{AA'}\Pi_{B'} = 0, \qquad Ⓗ$$

so $(\Omega^A, \Pi_{A'})$ are the spinor parts of a twistor (i.e. a constant local twistor).

For the "chopping and pasting" procedure, it is helpful to make a comparison with the electromagnetic case. Here, I refer to the magnetic charge that arises from Dirac's procedure which makes use of the non-global nature of the usual electromagnetic potential for a magnetic monopole. The relevant exact sequence is now (part of) the deRham sequence

$$0 \to \mathbb{C} \xrightarrow{i} \Omega^0 \xrightarrow{d} \Omega^1 \xrightarrow{d} \Omega^2 \qquad Ⓘ$$

(Ω^r being the sheaf of **holomorphic** r-forms) and the procedure amounts to evaluating a closed 2-form (the Maxwell field) on a 2-cycle which lies in an open region **R** (of homotopy type S^2) surrounding a world-tube of magnetic charge.



Breaking the sequence up as

$$0 \to d\Omega^0 \to \Omega^1 \to d\Omega^1 \to 0$$
$$0 \to \mathbb{C} \to \Omega^0 \to d\Omega^0 \to 0$$

(the relevant region being **R**, in each case), we get

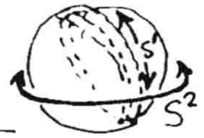
$$\curvearrowleft \; H^0(d\Omega^1) \to H^1(d\Omega^0) \to H^1(\Omega^1)$$

and

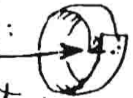
$$H^1(\Omega^0) \longrightarrow H^1(d\Omega^0) \to H^2(\mathbb{C}) \to H^2(\Omega^0).$$

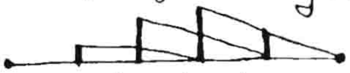
Since $H^1(\Omega^1) = H^1(\Omega^0) = H^2(\Omega^0) = 0$, and $H^2(\mathbb{C}) = \mathbb{C}$, we have a map from $H^0(d\Omega^1)$ onto $\mathbb{C}$, this being the evaluation of the total magnetic charge of the world-tube. Although this could have been obtained as a Gauss integral of the field over S^2, the above procedure is seen more explicitly in terms of the following ("chopping and pasting") operation. The Maxwell field F is a closed (whence locally exact) 2-form, and is therefore in $H^0(d\Omega^1)$. Think of $\mathbf{R}$ being chopped into two (slightly overlapping) hemispherical, topologically trivial, regions $\mathbf{R}_1$, $\mathbf{R}_2$. In each region $\mathbf{R}_i$, we have $F = d\overset{i}{A}$, whence

$$d(\overset{2}{A}-\overset{1}{A}) = d\overset{2}{A} - d\overset{1}{A} = F - F = 0$$



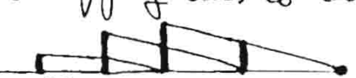
on the overlap (homotopy type S').

We can then chop this overlap into two — or just cut it once, with an overlap:  — and try to write $\overset{2}{A}-\overset{1}{A}$ as $d\varphi$. We find that the scalar field φ jumps as we go around, the difference between the two φ-values $\overset{\mathrm{I}}{\varphi}$, $\overset{\mathrm{II}}{\varphi}$ at the overlap being a constant (since $d(\overset{\mathrm{II}}{\varphi}-\overset{\mathrm{I}}{\varphi}) = d\overset{\mathrm{II}}{\varphi} - d\overset{\mathrm{I}}{\varphi} = (\overset{2}{A}-\overset{1}{A})-(\overset{2}{A}-\overset{1}{A}) = 0$), this constant being the required magnetic charge μ. In each case, we simply follow back along the de Rham sequence — which we can express graphically as the (Zeeman) diagram:



$$0 \to \{\mu\} \xrightarrow{i} \{\varphi\} \xrightarrow{d} \{A\} \xrightarrow{d} \{F\} \to 0 \qquad Ⓙ$$

We wish to apply this to the helicity 3/2 sequence Ⓐ



$$0 \to \left\{\begin{matrix}\Omega\\ \Pi\end{matrix}\right\} \to \left\{\begin{matrix}\omega\\ \pi\end{matrix}\right\} \to \left\{\begin{matrix}\rho\\ \sigma\end{matrix}\right\} \to \left\{\begin{matrix}0\\ \psi\end{matrix}\right\} \to 0$$

with a view to using it in curved Ricci-flat space-time $\mathcal{M}$. In $\mathcal{M}$, however, the space $\left\{\begin{matrix}\Omega\\ \Pi\end{matrix}\right\}$ does not exist (or becomes trivial) since the equations Ⓗ become inconsistent. Moreover, the equations Ⓑ and Ⓓ also become inconsistent, and must be abandoned. We can retain Ⓒ, Ⓔ, and Ⓕ, but in dropping Ⓑ, we can obtain an additional "field" quantity which lies in the extra gauge freedom

$$\rho_{A'BC} \mapsto \rho_{A'BC} + \overset{\bullet}{\rho}_{A'BC}$$

where

$$\nabla^{B}_{B'} \overset{\bullet}{\rho}_{A'BC} = 0 .$$

This "field" is a helicity $-3/2$ quantity $\tilde{\psi}_{ABC}\,(=\tilde{\psi}_{(ABC)})$, in flat space-time, defined by

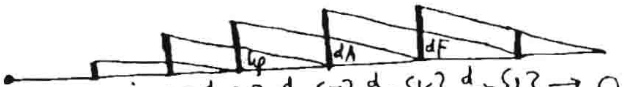
$$\nabla_A^{A'}\overset{\bullet}{\rho}_{A'BC}=\tilde{\psi}_{ABC}$$

(see LJM&RP, TN37), the gauge freedom in $\overset{\bullet}{\rho}$ being part of that (the "π-part") which is already a freedom in ρ. In $\mathbb{M}$, we still have an exact sequence

$$0\to\left\{{\Omega \atop \Pi}\right\}\to\left\{{\omega \atop \pi}\right\}\to\left\{{\rho \atop \sigma}\right\}\to\left\{{\tilde{\psi} \atop \psi}\right\}\to 0, \qquad ⓚ$$

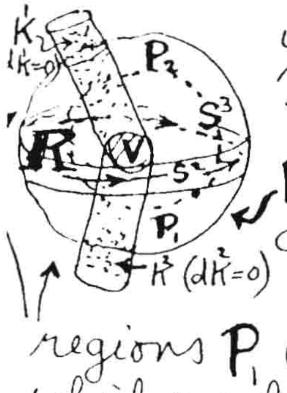
but now the twistor (Ω,Π) is not merely the charge Z^α for ψ, but contains a contribution from the projection part (i.e. secondary part) of the charge (dual twistor) $\mathbf{W}_\alpha$ for $\tilde{\psi}$. Thus, (Ω,Π) has the form "$Z^\alpha+I^{\alpha\beta}W_\beta$".

Of course, we require some sort of non-linear twistor space for $\mathcal{M}$, so it is just as well that a (Ω,Π) defined simply by Ⓗ will not do. For various suggestive (but, as yet, inadequate) reasons, it appears that it would be greatly helpful if the flat-space sequence ⓚ could be extended to a longer one. (Roughly: each "twistor (Ω,Π)" exists, in some sense, "somewhere" — say at $\mathbb{C}\mathscr{I}^+$ — but different twistors exist at different places, so they cannot be added. The vacuum equations allow this twistor concept to be "spread" about the space-time. However, ⓚ really "looks at" regions of $\mathcal{M}$ that are essentially no more than two-dimensional, and we need at least three dimensions to "feel out" the space-time.) The idea is that we try to mirror what goes on with the de Rham sequence Ⓘ extended further to the right, which I write as



$$0\to\{\mu\}\xrightarrow{i}\{\varphi\}\xrightarrow{d}\{A\}\xrightarrow{d}\{F\}\xrightarrow{d}\{K\}\xrightarrow{d}\{L\}\to 0\;.$$

Here, first, we allow the "Maxwell field" F to have a magnetic charge density K (a 3-form $K=dF$), so the standard Maxwell equations are violated. At this stage, this density is still conserved, however ($dK=0$, i.e. $\nabla_a{}^{*}K^a=0$), owing to the existence of F. But we can also allow some region **V** of violation of magnetic charge conservation, this charge violation being expressed by the 4-form L=dK, where K is not now of the form dF. Suppose that **V** is a small, topologically trivial, region of space-time, where we can imagine world tubes of conserved magnetic charge density entering and leaving **V**. Now, surround **V** by an (open) region **P** of space-time (homotopy type S^3). We can imagine **P** to be built up from two topologically trivial "hemispherical" regions $\mathbf{P}_1$ ("past hemisphere") and $\mathbf{P}_2$ ("future hemisphere") which overlap in a region **R** (homotopy type S^2) containing an "equatorial" S^2. In $\mathbf{P}_i$, we can find $\overset{i}{F}$ such that $d\overset{i}{F}=\overset{i}{K}$, for each i, and in **R** we have $F=\overset{2}{F}-\overset{1}{F}$, with $dF=0$.



From there on, we proceed as before, to obtain the magnetic charge μ — which now measures the total charge creation in the region **V**. More abstractly, we extend the original evaluation procedure to obtain a map from $H^0(\mathbf{P}, d\Omega^2)$ onto $H^3(\mathbf{P},\mathbb{C})\cong\mathbb{C}$. The field K, being locally of the form dF, is an element of $H^0(\mathbf{P}, d\Omega^2)$.

To extend the sequence Ⓚ, we make use of various ideas, and, most specifically, one due to Jörge Frauendiener (JF in TN 37). It will simplify matters if we first consider merely the projection part of Ⓚ suitably extended. In JF's scheme, instead of using a $\sigma_{A'B'C}$ subject to Ⓒ, we consider the weaker (conformally invariant) equation

$$\nabla^{B'}_{(B}\sigma_{C)A'B'}=0 \qquad \text{Ⓛ}$$

(with the symmetry $\sigma_{A'B'C}=\sigma_{(A'B')C}$ still holding), and then we define

$$\beta_{A'}=\nabla^{BB'}\sigma_{A'B'B} \qquad \text{Ⓜ}$$

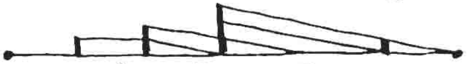
which, by virtue of the vacuum equations satisfies

$$\nabla^{AA'}\beta_{A'} = 0 \qquad Ⓝ$$

with $\beta_{A'}$ invariant under the same gauge freedom as before:

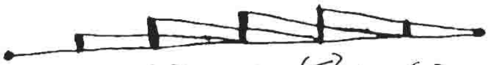
$$\dot{\sigma}_{A'B'C} \mapsto \dot{\sigma}_{A'B'C} + \nabla_{B'C}\pi_{A'}, \quad \nabla^{AA'}\pi_{A'} = 0. \qquad Ⓞ$$

However, the sequence is now not exact, and we have a Zeeman diagram looking like



$$0 \to \{\Pi\} \xrightarrow{i} \{\pi\} \xrightarrow{\nabla} \{\dot{\sigma}\} \xrightarrow{\nabla\cdot} \{\beta\} \to 0 \quad .$$

Exactness can be restored by introducing the somewhat odd-looking sheaf of "$\dot{\sigma}$ moddedout by π" to describe the "sourced spin $3/2$ field":



$$0 \to \{\Pi\} \to \{\pi\} \to \{\sigma\} \to \{\dot{\sigma}\}/\{\pi\} \to \{\beta\} \to 0 \qquad Ⓟ$$

and now β describes a kind of charge density for this field.

We can try to understand this by resorting to the abstract procedures adopted previously, where all homology groups refer to the region **P** described earlier, and we imagine some sort of "twistor creation" taking place in the region **V** that **P** surrounds. We have (in an obvious notation)

$$H^0(\dot{\sigma}/\pi) \to H^0(\beta) \to H^1(\nabla\sigma) \to H^1(\dot{\sigma}/\pi) \to H^1(\beta) \qquad Ⓠ$$

from

$$0 \to \{\nabla\sigma\} \to \{\dot{\sigma}\}/\{\pi\} \to \{\beta\} \to 0.$$

In $\mathbb{M}$, the full "twistor charge" for helicity $3/2$ should be found in $H^1(\nabla\sigma)$, since this contains the helicity raised "Gauss integral" for the field $\psi = \nabla\sigma$ defined by the potential σ according to Ⓓ. It can be seen that $H^1(\beta) = 0$, but we must examine $H^1(\dot{\sigma}/\pi)$. Using

$$0 \to \{\pi\}/\{\Pi\} \to \{\dot{\sigma}\} \to \{\dot{\sigma}\}/\{\pi\} \to 0$$

we obtain

$$H^1(\dot{\sigma}) \to H^1(\dot{\sigma}/\pi) \to H^2(\pi/\Pi) \to H^2(\dot{\sigma}).$$

By results due to JF and GAJS, it follows that $H^1(\dot{\sigma}) = H^2(\dot{\sigma}) = 0$, whence $H^1(\dot{\sigma}/\pi) \cong H^2(\pi/\Pi)$. Using

$$0\to\{\Pi\}\to\{\pi\}\to\{\pi\}/\{\Pi\}\to 0$$

we obtain $$H^2(\pi)\to H^2(\pi/\Pi)\to H^3(\Pi)\to H^3(\pi).$$
We have $H^2(\pi)=H^3(\pi)=0$, but $H^3(\Pi)$ is just the primed spin space $\mathbb{S}_{A'}$ $(\cong\mathbb{C}^2)$. This gives us back the "charge integral" that we would have obtained by directly applying a "chop and paste" procedure to σ. It delivers (in the absence of the second potential ρ) merely the projection part $\Pi_{A'}$ of the twistor $(\Omega^A,\Pi_{A'})$. However, the full twistor must be in $H^1(\nabla\sigma)$, whence it follows that the primary part lies in $H^0(\beta)$ (when $\Pi_{A'}=0$). Somehow, in the sequence Ⓟ, the twistor is "split" between the two ends of the sequence, Ω^A being found in H^0 of the right-hand end and $\Pi_{A'}$ in H^3 of the left-hand end.

To proceed further, we must re-introduce the second potential ρ. This gives the whole scheme another (short) exact sequence structure, involving
$$0\to\{\mathring{\dot\rho}\}\to\{\dot\rho\}\to\{\dot\sigma\}\to 0,$$
where Ⓛ must be filled out to
$$\nabla^{B'}_{(B}\dot\sigma_{C)A'B'}=0,\quad \nabla^{B}_{(B'}\dot\rho_{A')BC}=2i\dot\sigma_{A'B'C} \qquad \text{Ⓡ}$$

$\mathring{\dot\rho}_{A'BC}$ subject to $\nabla^{B}_{(B'}\mathring{\dot\rho}_{A')BC}=0$ acting as freedom in $\dot\rho$ (each spinor symmetric in the relevant index pair). All this works in Ricci-flat $\mathcal{M}$, with gauge freedom for $\dot\rho,\dot\sigma$ the same as with ρ,σ, as before, in Ⓔ, Ⓕ. In $\mathbb{M}$, we get

$$\begin{tikzcd}
 & 0 \arrow[d, dashed] & 0 \arrow[d] & 0 \arrow[d, dashed] & 0 \arrow[d] & 0 \arrow[d, dashed] & \\
0 \arrow[r] & \{\mathring\Omega\} \arrow[r] \arrow[d] & \{\mathring\omega\} \arrow[r] \arrow[d] & \{\mathring\rho\} \arrow[r] \arrow[d] & \{\mathring{\dot\rho}\}/\{\mathring\omega\} \arrow[r] \arrow[d] & \{\mathring\alpha\} \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & \left\{\begin{smallmatrix}\Omega\\ \Pi\end{smallmatrix}\right\} \arrow[r] \arrow[d] & \left\{\begin{smallmatrix}\omega\\ \pi\end{smallmatrix}\right\} \arrow[r] \arrow[d] & \left\{\begin{smallmatrix}\rho\\ \sigma\end{smallmatrix}\right\} \arrow[r] \arrow[d] & \left\{\begin{smallmatrix}\dot\rho\\ \dot\sigma\end{smallmatrix}\right\}/\left\{\begin{smallmatrix}\omega\\ \pi\end{smallmatrix}\right\} \arrow[r] \arrow[d] & \left\{\begin{smallmatrix}\alpha\\ \beta\end{smallmatrix}\right\} \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & \{\Pi\} \arrow[r] \arrow[d] & \{\pi\} \arrow[r] \arrow[d] & \{\sigma\} \arrow[r] \arrow[d] & \{\dot\sigma\}/\{\pi\} \arrow[r] \arrow[d] & \{\beta\} \arrow[r] \arrow[d] & 0 \\
 & 0 & 0 & 0 & 0 & 0 &
\end{tikzcd} \qquad \text{Ⓢ}$$

as a commutative exact diagram of sheaves. Here, we have

$$\alpha_C = \tfrac{1}{2}\nabla^{BB'}\rho_{B'BC}\,,\quad \beta_{A'} = \nabla^{BB'}{\odot}\sigma_{A'B'B} \qquad \text{Ⓣ}$$

(the "freedom" in α and ρ being denoted $\mathring{\alpha}$, $\mathring{\rho}$, etc.) and the equations

$$\nabla^{A}_{A'}\alpha_A = 2i\,\beta_{A'}\,,\quad \nabla^{A'}_{A}\beta_{A'} = 0$$

(compare Ⓕ). As sheaf sequences, everything seems to work consistently from the third ($\{\omega\}$, $\{\substack{\omega\\ \pi}\}$, $\{\pi\}$) column out to the right, provided that $\mathcal{M}$ is Ricci-flat. In $\mathbb{M}$, we have, for the second column, the standard Poincaré invariant twistor short exact sequence

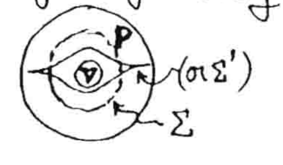
$$0 \rightarrow \mathbb{S}^A \rightarrow \mathbb{T}^\alpha \rightarrow \mathbb{S}_{A'} \rightarrow 0.$$

In fact, we find four versions of this sequence (two in dual form) running round the outside of the square

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 & \mathbb{S}_A \arrow[l] \arrow[d] & \mathbb{T}_\alpha \arrow[l] \arrow[d] & \mathbb{S}^A \arrow[l] \arrow[d] & 0 \arrow[l] \\
0 & \mathbb{T}^\alpha \arrow[l] \arrow[d] & \boxed{\begin{array}{l}\text{Charges}\\ \text{for } \psi\\ \text{and } \varphi\end{array}} \arrow[l] \arrow[d] & \mathbb{T}_\alpha \arrow[l] \arrow[d] & 0 \arrow[l] \\
0 & \mathbb{S}_{A'} \arrow[l] \arrow[d] & \mathbb{T}^\alpha \arrow[l] \arrow[d] & \mathbb{S}^A \arrow[l] \arrow[d] & 0 \arrow[l] \\
 & 0 & 0 & 0 &
\end{tikzcd}$$

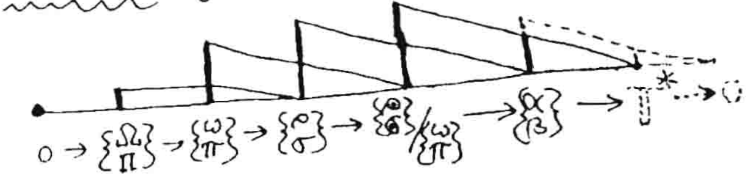
There is a certain "twist" in this diagram owing to the "$Z^\alpha + I^{\alpha\beta}W_\beta$" nature of (Ω, Π), as mentioned earlier (after Ⓚ). The second column of this sequence comes from H^3 of the second column of Ⓢ, and the fourth column comes from H^0 of the sixth column of Ⓢ.

We can also introduce duals for the spaces of these various types of "field" in the region $\mathbf{P}$, constructing a divergence-free vector J^a, to be integrated over some topologically non-trivial compact 3-surface Σ ($\cong S^3$) within $\mathbf{P}$. Remarkably, we find that the relevant dual sheaf sequence is simply the same sequence as Ⓟ all

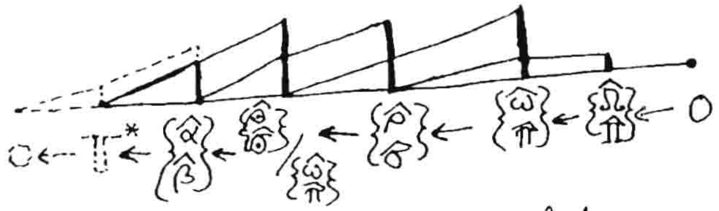


over again, but running in the opposite direction:

Dual of


$0 \to \begin{Bmatrix}\Omega\\ \Pi\end{Bmatrix} \to \begin{Bmatrix}\omega\\ \pi\end{Bmatrix} \to \begin{Bmatrix}\rho\\ \sigma\end{Bmatrix} \to \begin{Bmatrix}\overset{\odot}{\rho}\\ \overset{\odot}{\sigma}\end{Bmatrix} \Big/ \begin{Bmatrix}\omega\\ \pi\end{Bmatrix} \to \begin{Bmatrix}\alpha\\ \beta\end{Bmatrix} \to \mathbb{T}^* \dashrightarrow 0$

is


$0 \dashleftarrow \mathbb{T}^* \leftarrow \begin{Bmatrix}\hat\alpha\\ \hat\beta\end{Bmatrix} \leftarrow \begin{Bmatrix}\hat{\overset{\odot}{\rho}}\\ \hat{\overset{\odot}{\sigma}}\end{Bmatrix} \Big/ \begin{Bmatrix}\hat\omega\\ \hat\pi\end{Bmatrix} \leftarrow \begin{Bmatrix}\hat\rho\\ \hat\sigma\end{Bmatrix} \leftarrow \begin{Bmatrix}\hat\omega\\ \hat\pi\end{Bmatrix} \leftarrow \begin{Bmatrix}\hat\Omega\\ \hat\Pi\end{Bmatrix} \leftarrow 0$

where to express this duality we use

$$J^{AA'} = \omega^A \hat\beta^{A'} + \pi^{A'} \hat\alpha^A \quad \text{or} \quad \rho^{AB}_{B'} \hat{\overset{\odot}{\sigma}}{}^{A'B'}_{B} + \sigma^{A'B'}_{B} \hat{\overset{\odot}{\rho}}{}^{AB}_{B'}$$

$$\text{or} \quad \overset{\odot}{\rho}{}^{AB}_{B'} \hat\sigma^{A'B'}_{B} + \overset{\odot}{\sigma}{}^{A'B'}_{B} \hat\rho^{AB}_{B'} \quad \text{or} \quad \alpha^A \hat\pi^{A'} + \beta^{A'} \hat\omega^A,$$

as the case may be (possibly modulo factors). The dotted part at the end of each sequence is there if we consider fields throughout **P** (i.e. $H^0(\mathbf{P},\ )$), whereas the dotted part is absent if we merely consider sheaves. The existence of this duality serves to confirm some of the statements concerning $H^0(\mathbf{R}, \beta)$, etc., made earlier.

There is still a good deal that remains mysterious about what happens in a general Ricci-flat M. Work in progress.

Thanks to LJM and JF (and also GAJS) especially.

Roger P

References

R. Penrose (1992) Twistors as spin 3/2 charges, in Gravitation and Modern Cosmology. eds. A. Zichichi, N. de Sabbata, and N. Sánchez (Plenum Press, New York)

R. Penrose (1994) Twistors and the Einstein Equations, in proceedings of Devon conference at Seale Hayne, ed. S.A. Huggett (to appear)

234

ON THE SECOND LAW OF THERMODYNAMICS

J. Statist. Phys. **77** (1994), no. 1–2, 217–21

In honour of my brother Oliver, where I contrast our different concerns, in relation to the ubiquitous Second Law.

On the Second Law of Thermodynamics

Roger Penrose[1]

Received February 16, 1994

The second law of thermodynamics has two distinct aspects to its foundations. The first concerns the question of why entropy goes up in the future, and the second, of why it goes down in the past. Statistical physicists tend to be more concerned with the first question and with careful considerations of definition and mathematical detail. The second question is of quite a different nature; it leads into areas of cosmology and quantum gravity, where the mathematical and physical issues are ill understood.

KEY WORDS: Second law of thermodynamics; time asymmetry; Weyl curvature hypothesis.

TWO ASPECTS OF THE SECOND LAW

The very brief remarks that I am making here can in no way reflect the magnitude of the scientific and mathematical indebtedness that I feel towards my brother Oliver Penrose. From a very young age, he provided me both with inspiration and with a depth of realization that in mathematics and in physics there was to be found excitement, mystery, and much logical structure of profound beauty. Though at that early stage in life I had imagined that I would become a doctor, there was always that older brother whom I greatly looked up to, and from whom I gained a profound appreciation that in the "hard sciences" also there was something to strive for that is deeply worthwhile. Only much later, after my early flirtations with pure mathematics, did I finally turn to the study of the physical world in any serious way. The many conversations with Oliver, over a great number of influential years, have helped me to formulate my viewpoints and sharpen my critical sense.

Oliver's own interests took him in the direction of statistical physics,

---

[1] Mathematical Institute, Oxford, OX1 3LB, United Kingdom.

where his profound contributions are acknowledged world-wide. Although I had picked up, from Oliver, some of the underlying concepts, statistical physics is an area in which I have never myself felt comfortable; so it was with some considerable awkwardness and uncertainty that I have found myself driven—as a logical consequence of my own interests in the structure of black holes and cosmological space-time singularities—to enter into the discussion of time asymmetry and the origin and nature of the second law of thermodynamics. It seems that coming at this subject as an outsider with an unconventional point of view, I have found myself to be intrigued by issues that appear to be different from those that have traditionally been the concern of most statistical physicists. Sometimes such people would address me with some puzzlement at the fact that my interests in this area would be so much concentrated on matters of cosmology and quantum gravity, while I would seem almost totally to ignore that great body of literature on such matters as Boltzmann's H theorem, microcanonical ensembles, equilibrium fluctuations, and the like. Statistical physicists are themselves not usually particularly interested (professionally) in cosmology —so I might be informed—and certainly not in quantum gravity. Those subjects, it would appear, have almost no relevance to the matters of primary concern to a statistical physicist. On the issues that might more directly *be* their concern, my own silence would, more often than not, merely reflect my own considerable lack of expertise; nevertheless I feel that there is a separate question of relevance here, and it is this that I shall try to address in this short note.

As I see it, there is at least one important difference in motivation between my own interests and those of the main body of statistical physicists. Most conventional physicists are concerned with predicting the future, rather than retrodicting the past. To them, if the second law of thermodynamics holds any mystery—or even if it merely falls marginally short of being rigorously understood—what would need to be understood would be: why it is that the entropy in an isolated system goes *up* in the *future*? To me, the mystery is a different one: why does the entropy go *down* in the *past*?

Superficially, these might seem to be the same question. But on reflection, one sees that the two questions are completely different. We may recall that Boltzmann himself(1) had promoted the idea that the existence of a second law of thermodynamics at our present epoch might be the result of our universe—or at least of our local part of the universe—having encountered an enormous fluctuation of low entropy from which our present increasing entropy would represent a natural proceeding toward thermal equilibrium. On this view, at the times that immediately precede that fluctuation, the time direction of entropy increase would have been opposite to what we experience now. In this picture of things, the entropy of the universe would go up into the past (that is, in the past direction, to the past of this fluctuation) just as it goes up, now, in the time direction that we now refer to as the "future." All the reasoning that leads us to expect that the entropy of our present universe (or universe neighborhood) is increasing into the future, away from this fluctuation, would apply equally well in the past direction, to the past of this fluctuation. In such a scheme, the second law would *not* have held in the past, with respect to the time direction that we use *now*. However, all the cosmological evidence points overwhelmingly to the conclusion that there has been no such turnaround in the direction of the second law—at least so far as our direct (and indirect) observations of the early universe imply. If our second law is the result of an enormous fluctuation, then it must have been a fluctuation of much *greater* enormity and at a much earlier time than Boltzmann could have imagined. Indeed the fluctuation would have to have been so enormous that no explanation of the "anthropic" kind—whereby the very existence of intelligent life provides the reason for the low-entropy "oasis" from which we need to have emerged (cf. ref. 2, p. 354; also ref. 3, p. 589)—comes close to providing an explanation. Thus, the fact that the entropy continues to go *down* in the past, the farther into the past we probe, provides us with a mystery of a quite different order from the more familiar problem of showing why it is that the entropy has a tendency to increase in the future.

These two aspects of the second law, it seems to me, have a quite distinct conceptual status, and the types of mathematical and physical argument that are brought to bear on them appear to be very different. For those arguments that are aimed at showing why the entropy goes up in the future— or, at least, *usually* goes up in the future—one must have a deep understanding of the meaning of the term "entropy" and of the precise mathematical formulations of the various concepts involved. One must also be explicit about the assumptions that need to be invoked in the mathematical derivations. (See Oliver's classic text(4) for a rigorous and comprehensive treatment of these issues.) To ensure that there is a general entropy increase into the future, one must assume the absence of prior correlations in the details of the initial particle motions. It seems "reasonable" to make such an assumption, since within any chosen set of macroscopic parameters, those initial states for which such correlations are essentially absent would form the vast majority. Nevertheless, one must expect that correlations of this nature would be *present*, in *time-reversed* sense, *after* any significant increase in entropy, so one cannot assume, just on general grounds, that the absence of correlations is an all-embracing feature of our inverse. In an important paper, Oliver, together with Ian

Percival,(5) set down an explicit assumption that provides a past/future distinction, which they referred to as the *law of conditional independence*. According to this assumption, the absence of such correlations would be a feature of the remote past, so that deductions concerning the general increase in entropy in the future can then be made (contrast ref. 3, p.589).

My own interest in the second law, on the other hand, has been concerned with the reverse question of why the entropy goes *down* in the *past*. Here, the effect is gross rather than subtle. Delicate issues concerning the definition of entropy, the nature of equilibrium and of fluctuations away from equilibrium, etc.—though they are certainly of great importance when one is concerned with details of the *future* behavior of an isolated system—appear to lie far from the gross type of explanation that would seem to be needed for an understanding of this particular aspect of the second law. My own attempts at an explanation have been described in a series of papers over the years,(6, 3) where I have been promoting a very different kind of criterion for the initial state of the universe from the law of conditional independence of Penrose and Percival, which I have referred to as the *Weyl curvature hypothesis*. It has very little explicit connection with the notion of entropy, as such. Even my earlier attempts at finding a general definition of "gravitational entropy" have had to be abandoned—or at least set aside, for the time being. It is only the Bekenstein–Hawking formula for black-hole entropy which has so far found any specific role to play in my deliberations, namely to provide an estimate for the total entropy of a closed universe. (One finds, by use of this formula, that the entropy in a generic closed universe with baryon number 10^{80} is about 10^{123}, in natural units.)

The role of the Weyl curvature hypothesis is to set up an initial state for the universe which is very special in a particular way—a way that is consistent with what we know about cosmology and astronomy. That the entropy of the early universe is thereby constrained to be extremely low—incomparison with what it would otherwise have been (recall the figure 10^{123})—is, in a sense, incidental to this. But the hypothesis does indeed provide a picuture in which the entropy must go down in the past. On the other hand, it provides no guarantee that the entropy will continue to go up in the future. Perhaps some kind of combination of the law of conditional independence with the Weyl curvature hypothesis would be needed if an understanding of the second law of thermodynamics, in *both* these aspects, is to be obtained (cf. also ref. 3, pp. 633–634).

It may be remarked that, although the Weyl curvature hypothesis may be thought of as a "blunt instrument" in its relation to thermodynamics, it has played a clear-cut role in some precise mathematical theorems of importance to cosmology. In particular, Newman(7) has shown, using a refined form of the Weyl curvature hypothesis due to K. P. Tod—whereby it is assumed that the initial big-bang singularity is conformally regular as a hypersurface—that one can deduce that, with appropriate equations of state, the early universe must be of Friedmann–Robertson–Walker type (as, indeed, it is observed closely to be).

The other main role of the Weyl curvature hypothesis lies in highly speculative matters to do with quantum gravity and with the measurement problem in quantum mechanics. My own interests have taken me in the direction of these particular questions,(8) but it will be a long while before the mathematical status of any of these considerations comes close to that enjoyed within the subject of statistical physics.

REFERENCES

1. L. Boltzmann, *Nature* **51**:413 (1895).
2. R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, Oxford, 1989).
3. R. Penrose, Singularities and time-asymmetry, in *General Relativity: An Einstein Centenary*, S. W. Hawking and W. Israel, eds. (Cambridge University Press, Cambridge, 1979).
4. O. Penrose, *Foundations of Statistical Mechanics: A Deductive Treatment* (Pergamon Press, Oxford, 1970).
5. O. Penrose and I. C. Percival, *Proc. Phys. Soc.* **79**:509 (1962).
6. R. Penrose Singularities in cosmology, in *Confrontation of Cosmological Theories with Observational Data*, M. S. Longair, ed. (Reidel, Boston, 1974); Singularities of space-time, in *Theoretical Principles in Astrophysics and Relativity*, N. R. Liebowitz, W. H. Reid, and P. O. Vandervoort, eds. (Chicago University Press, Chicago, 1978).
7. R. P. A. C. Newman, On the structure of conformal singularities in classical general relativity, *Proc. R. Soc. A* **443**:473–492 (1993); II, Evolution equations and a conjecture of K. P. Tod, *Proc. R. Soc. A* **443**:517–546 (1993).
8. R. Penrose, Time-asymmetry and quantum gravity, in *Quantum Gravity 2*, D. W. Sciama, R. Penrose, and C. J. Isham, eds. (Oxford University Press, Oxford, 1981); Gravity and statevector reduction, in *Quantum Concepts in Space and Time*, R. Penrose and C. J. Isham, eds., (Oxford University Press, Oxford, 1986), pp. 126–146; Gravity and quantum mechanics, in *General Relativity and Gravitation 1992*, R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi, eds. (Institute of Physics, Bristol, 1993).

235

TWISTORS AND THE EINSTEIN EQUATIONS

Twistor Theory. Ed. S. Huggett, 145–57. *Lecture Notes in Pure and Appl. Math.*, 169, Dekker, New York, 1994

492 Twistors and the Einstein equations. *Twistor theory.* Ed. S. Huggett, 145–57. *Lecture Notes in Pure and Appl. Math.*, 169, 1994

11

Twistors and the Einstein Equations

Roger Penrose
Mathematical Institute
Oxford, OX1 3LB, U.K.

Abstract It has been proposed that the appropriate global definition of a twistor, applicable to general curved vacuum space-times, would be as a charge for a massless field of helicity $^3/_2$. In flat space-time, using the Dirac form of these potentials, these twistor charges arise as the "gauge freedom of the second kind" in a long exact sequence involving the first and second potentials for the field.

A construction due to Ward is recalled, in which potentials for massless fields can act as partial connections on non-linear bundles, integrable on β-planes. This is generalized, in the case of helicity $^3/_2$, to provide a full connection on a vector bundle of rank 3, leading to an expression whereby the usual Rarita-Schwinger potential is supplemented by a second potential.

1. Twistors in curved space-time

From the very beginning, twistor theory has been motivated by the desire for a holomorphic underpinning to space-time structure (cf. Penrose 1986). When (null projective) twistors are identified with light rays in space-time, this twistor holomorphicity shows up in the conformal structure of the celestial sphere and in the shear-free property of families of light rays. Yet general relativity has a basic requirement that (vacuum) space-times be conformally curved, in general. This leads to an essential incompatibility with the foregoing manifestations of twistor holomorphicity, because the shear-free condition is not generally preserved along light rays in the presence of conformal curvature and, accordingly, the conformal structures of the local and asymptotic celestial spheres disagree with one another. It might thus *appear* to be the case that there is a deep-seated incompatibility between twistor-holomorphic requirements and the principles of general relativity.

If this were really so, it would, of course, be "bad news" for twistor theory's aspirations to occupy significant place in the basic formulation of Nature's laws — especially since Einstein's vacuum equations are now observationally confirmed to a degree greater than that of any other physical theory. But despite these initially unpromising-looking conflicts, there have emerged, from time to time, some suggestive — though as yet somewhat limited — relationships between twistor holomorphicity and the specific form of the Einstein vacuum equations. In the first place, the Hamiltonian that governs the transformation between the flat twistor spaces on the two sides of a plane or spherically fronted gravitational impulsive wave (as is determined by the way in which the wave deflects light rays) is specified in terms of a holomorphic function (Penrose and MacCallum 1972), this being a characteristic feature of the Einstein equations holding along the wave. More strikingly, the $\mathcal{H}$-space and non-linear graviton constructions (Newman 1976, Penrose 1976, 1992a) provide direct links between twistor holomorphicity and the Einstein vacuum equations, albeit only in the case of (anti-) self-dual Weyl curvature. These are perhaps the only unambiguous manifestations of twistor holomorphicity in the non-linear Einstein vacuum equations, although in the weak field linear limit of Einstein's equations, twistor holomorphicity provides a complete description through the standard contour integral and cohomology expressions (cf. Penrose 1969, Penrose and Rindler 1986, Eastwood, Penrose, and Wells 1981). As is well known, the non-linear graviton provides a successful "non-linearization" of the helicity -2, or anti-self-dual case (twistor function homogeneity +2); moreover the "googly" construction provides a very tentative "non-linearization" of the helicity +2, or self-dual case (twistor function homogeneity -6). The ambitwistor approach to the full Einstein vacuum equations has had some success within its own terms (LeBrun 1990). Hypersurface twistors provide a means that can in principle lead to a formulation of the full vacuum equations in twistor terms (Mason 1985, Mason and Penrose 1989, cf. also Penrose 1990a). The Ward-Woodhouse-Mason construction for axisymmetric stationary vacuums (Ward 1983, Woodhouse and Mason 1988) provides a non-standard route to a twistor-holomorphic description of an interesting family of solutions, but it is not clear how this might be generalized to handle general vacuum fields.

Indeed, it is not clear from *any* of this how one might envisage a twistor concept that applies to Einstein vacuum equations in complete generality. It is for this kind of reason that we may feel encouraged by the existence of a another approach to the twistorial description of the Einstein equations.

2. Twistors as charges for helicity $^3/_2$ massless fields

In Penrose (1990b, 1991, 1992b, and 1993) the suggestion was explored that one should combine the following two facts concerning massless fields of helicity $^3/_2$:

(i) the appropriate field equations governing massless fields of helicity $^3/_2$ are consistent if and only if the space-time is Ricci-flat;

(ii) in Minkowski space-time $\mathbb{M}$, the space of conserved charges for massless fields of helicity $^3/_2$ is its twistor space $\mathbb{T}$.

In accordance with (i) and (ii), it would seem that all we need do, in order to construct the elusive twistor space $\mathcal{T}$ for a general Ricci-flat space-time $\mathcal{M}$, is to find the space of charges for massless fields of helicity $^3/_2$ on $\mathcal{M}$. The hope would be that, as was the case for the non-linear graviton construction, the complex space $\mathcal{T}$ would be completely equivalent to the space-time $\mathcal{M}$, although related to it only in a non-local way. One might also hope that the general such $\mathcal{T}$ could be constructed by the use of free functions, and that twistor theory would thereby establish itself as providing a general procedure for analysing the Einstein vacuum equations. If these aims could be realized, then there would be a powerful case for considering that the physically "correct" twistor concept had at last been found.

However, it turns out that there are several profound difficulties in implementing this idea (see, in particular, Penrose 1992b). An over-riding obstacle to finding a "*T*", according to the above ideas is that it ought *not* to be a vector space (as is the case with the non-linear graviton), whereas in any of the normal definitions of a "charge" for a linear field would indeed provide us with a vector-space structure. The most promising way of circumventing this problem would seem to be to find way in which helicity $^3/_2$ fields can play some kind of active role as a *connection* (cf. Penrose 1991), in addition to its passive role in which the twistor charge acts as its source. Combining such active and passive roles might give rise to the desired non-linearity and thus break the unwanted vector-space structure.

In the present note, I shall consider both the passive and active aspects of fields of helicity $^3/_2$. However, much of what I shall have to say will be concerned with such fields in Minkowski space-time $\mathbb{M}$.

3. The Dirac and Rarita-Schwinger formulations

In $\mathbb{M}$, a field of helicity $^3/_2$ is described (using the 2-component spinor notation, cf. Penrose and Rindler 1984), as a solution of the equation

$$\nabla^{AA'}\varphi_{A'B'C'} = 0,$$

where

$$\varphi_{A'B'C'} = \varphi_{(A'B'C')},$$

with round brackets denoting symmetrization in the usual way. In a general space-time, these equations are inconsistent because of Buchdahl conditions relating $\varphi_{A'B'C'}$ to the (self-dual) conformal curvature $\Psi_{A'B'C'D'}$:

$$\Psi_{A'B'C'D'}\,\varphi^{A'B'C'} = 0.$$

For this reason, a potential $\sigma^{A}{}_{B'C'}$ (modulo a gauge freedom) is employed, instead of a gauge-invariant field.

In *flat* space-time $\mathbb{M}$, two such descriptions have been given, namely that of Dirac (1936) and of Rarita and Schwinger (1941). According to the Dirac formulation, we have

$$\nabla_{AA'}\sigma^{A}{}_{B'C'} = \varphi_{A'B'C'} \quad \text{and} \quad \nabla^{BB'}\sigma^{A}{}_{B'C'} = 0,$$

where $\sigma^{A}{}_{B'C'}$ is symmetric in and B' and C'. The gauge freedom for $\sigma^{A}{}_{B'C'}$ is

$$\sigma^{A}{}_{B'C'} \mapsto \sigma^{A}{}_{B'C'} + \nabla^{A}{}_{B'}\nu_{C'}$$

for which $\nu_{A'}$ satisfies the Weyl (anti-)neutrino equation

$$\nabla^{AB'}\nu_{B'} = 0.$$

In the Rarita-Schwinger form (in 2-component spinor notation), no symmetry condition is imposed on $\sigma^{A}{}_{B'C'}$, and we have the pair of field equations

$$\nabla^{B'(C}\sigma^{A)}{}_{B'C'} = 0, \quad \nabla_{C(C'}\sigma^{C}{}_{A')}{}^{C'} = 0,$$

with the same gauge freedom as above, except that $\nu_{A'}$ is not now restricted to satisfy any equation. (The Dirac description is equivalent to that of Rarita-Schwinger, in a special choice of gauge.) The second of the above equations tells us that $\nabla_{C(C'}\sigma^{C}{}_{A')B'}$ is totally symmetric in its primed indices, and we can now define the gauge-invariant field $\varphi_{A'B'C'}$ by

$$\nabla_{C(C'}\sigma^{C}{}_{A')B'} = \varphi_{A'B'C'}.$$

In both the Dirac and Rarita-Schwinger formulations, the equations for the potential and the gauge freedom are still consistent in *curved* space-time $\mathcal{M}$, if (and only if) $\mathcal{M}$ is Ricci-flat. However, in each case, the "field" $\varphi_{A'B'C'}$, as defined above, is not gauge invariant. Thus, one must be content with a "potential-modulo-gauge" description. The fact that the Ricci-flat condition is necessary and sufficient for the consistency of the gauge freedom follows immediately from the spinor form of the Ricci identities (as given in Penrose and Rindler 1984). The consistency of the field equations on $\sigma^{A}{}_{B'C'}$ also follows from the Ricci-flat condition (see, for example, Buchdahl 1958, Deser and Zumino 1976, Julia 1982).

Twistors and the Einstein equations. *Twistor theory.* Ed. S. Huggett, 145–57. *Lecture Notes in Pure and Appl. Math.*, 169, 1994

4. Twistors as gauge freedom of the second kind

In flat space-time $\mathbb{M}$, one can see the role of twistors as "charges" for massless fields of helicity $3/2$ very succinctly (in the Dirac form of the field equations) when a "second potential" is introduced, and then twistors appear as providing the "gauge freedom of the second kind". The Dirac-type second potential would be a quantity $\rho^{AB}{}_{C'}$, symmetric in A and B, where

$$\nabla_{BB'}\rho^{AB}{}_{C'} = \sigma^{A}{}_{B'C'} \quad \text{and} \quad \nabla^{CC'}\rho^{AB}{}_{C'} = 0.$$

The extra gauge freedom in $\rho^{AB}{}_{C'}$, over and above that in $\sigma^{A}{}_{B'C'}$ is given by

$$\rho^{AB}{}_{C'} \mapsto \rho^{AB}{}_{C'} + \nabla^{B}{}_{C'}\mu^{A},$$

where μ^A satisfies the Weyl neutrino equation

$$\nabla_{AC'}\mu^{A} = 0.$$

Combining the two gauge freedoms together, we find that

$$\rho^{AB}{}_{C'} \mapsto \rho^{AB}{}_{C'} + \varepsilon^{AB}\nu_{A'} + \nabla^{B}{}_{C'}\mu^{A},$$

and instead of satisfying the free Weyl neutrino equation, μ^A must have the gauge field $\nu_{B'}$ (for $\sigma^{A}{}_{B'C'}$) as a "source":

$$\nabla_{AC'}\mu^{A} = 2\nu_{C'}$$

(from the symmetry of $\rho^{AB}{}_{C'}$).

The *gauge freedom of the second kind* is (as with standard electromagnetism) that freedom in the gauge fields which does not affect the potentials. Thus, this freedom, in the present situation, is given when $\nu_{C'}$ is a constant and is related to μ^A by

$$\varepsilon^{AB}\nu_{A'} + \nabla^{B}{}_{C'}\mu^{A} = 0,$$

that is to say, the pair $(\omega^A, \pi_{A'})$ is a *twistor*, where

$$\omega^A = \mu^A, \quad \pi_{A'} = i\,\nu_{A'}.$$

The various relationships between these gauge quantities fit into an exact sequence:

$$0 \rightarrow \{\text{twistors}\} \rightarrow \{\text{gauge fields}\} \rightarrow \{\text{potentials}\} \rightarrow \{\text{fields}\} \rightarrow 0$$

where "{twistors}" stands for the space of pairs $(\omega^A,\pi_{A'})$, related by $\nabla_{AA'}\omega^B=-i\varepsilon_A{}^B\pi_{A'}$; "{gauge fields}" for the space of pairs $(\mu^A,\nu_{C'})$, related by $\nabla_{AC'}\mu^A=2\nu_{C'}$, $\nabla^{AC'}\nu_{C'}=0$; "{potentials}" for the space of $(\rho^{AB}{}_{C'},\sigma^A{}_{B'C'})$, each symmetric and related by $\nabla_{BB'}\rho^{AB}{}_{C'}=\sigma^A{}_{B'C'}$, $\nabla^{CC'}\rho^{AB}{}_{C'}=0$, $\nabla^{BB'}\sigma^A{}_{B'C'}=0$; and "{fields}" by the space of symmetric $\varphi_{A'B'C'}$, subject to $\nabla^{AA'}\varphi_{A'B'C'}=0$. The role of the twistors as "charges" for the fields $\varphi_{A'B'C'}$ can then be understood in terms of the cohomology of this sequence (cf. Penrose 1992b).

5. The Ward non-linear connection

In an early Twistor Newsletter (see Ward 1979) Richard Ward showed how to encode the information of a helicity s ($>{}^1/_2$) massless field into a non-linear partial connection on the (unprimed) spin-bundle over $\mathbb{CM}$ (complexified Minkowski space) which is integrable over β-planes. Let λ_A be the "fibre coordinate", and define a partial connection by the operator

$$\mathbf{W}_{A'} = \lambda^A \nabla_{AA'} - \rho_{A'}\boldsymbol{\Upsilon},$$

where $\boldsymbol{\Upsilon}$ is the Euler operator along the fibres:

$$\boldsymbol{\Upsilon} = \lambda^A\, \partial/\partial\lambda^A.$$

Note that the operator $\mathbf{W}_{A'}$ is defined only in the tangent directions to the β-plane defined by λ_A at the point in question. We require that $\mathbf{W}_{A'}$ be integrable on each β-plane (as defined by λ_A), and this condition amounts to

$$\lambda^A \nabla_{AA'}\rho^{A'} = \rho_{A'}\boldsymbol{\Upsilon}\rho^{A'},$$

where $\tau_{A'}$ is defined on the bundle — so it depends on λ_A as well as on the space-time point. Let us take $\rho_{A'}$ to have homogeneity degree m in λ_A and to have the polynomial form

$$\rho_{A'} = \lambda_B \ldots \lambda_D \, \rho_{A'}{}^{B \ldots D},$$

where $\rho_{A'}{}^{B \ldots D}$ (symmetric in the m indices B,..., D) is now a field depending only on the space-time point. Then the condition of integrability on β-planes becomes

$$\nabla^{A'(A} \rho_{A'}{}^{B \ldots D)} = 0.$$

This last relation is the equation for a (conformally invariant) penultimate potential for the massless field $\varphi_{A'B' \ldots D'}$ of helicity $s=(m+1)/2$, the field $\varphi_{A'B' \ldots D'}$ (symmetric in its $m+1$ indices) being defined by

$$\varphi_{A'B' \ldots D'} = \nabla^B{}_{(B'} \ldots \nabla^D{}_{D'} \rho_{A')B \ldots D},$$

and consequently satisfying

$$\nabla^{AA'} \varphi_{A'B' \ldots D'} = 0.$$

Apart from the final remarks concerning the field $\varphi_{A'B' \ldots D'}$, all these considerations apply in a conformally self-dual complex space-time as well as in flat space-time. A version of the expression for $\varphi_{A'B' \ldots D'}$ holds also if the space-time is conformally flat.

6. The case of helicity $^3/_2$

The case relevant to our present situation is given when m=2, and we have

$$W_{A'} = \lambda^A \nabla_{AA'} - \lambda_B \lambda_C \, \rho_{A'}{}^{BC} \lambda^A \, \partial / \partial \lambda^A$$

If we consider this as a differential operator acting on spinor fields λ_D defined on the space-time, we can write this

$$\mathbf{W}_{A'}\,\lambda_D = \lambda^A\,\nabla_{AA'}\,\lambda_D - \lambda_B\lambda_C\,\rho_{A'}{}^{BC}\,\lambda_D.$$

Setting

$$\mathbf{W}_{A'} = \lambda^A\,\Omega_{AA'},$$

where the operator $\Omega_{AA'}$ acts on spinor fields λ_D, we can write

$$\lambda^A\,\Omega_{AA'} = \lambda^A\,\nabla_{AA'} - \lambda_B\lambda_C\,\rho_{A'}{}^{BC}.$$

Integrability on β-planes yields

$$\nabla^{A'(A}\rho_{A'}{}^{BC)} = 0,$$

and in flat space-time, we have the *field*

$$\varphi_{A'B'C'} = \nabla^B{}_{(B'}\nabla^C{}_{C'}\rho_{A')BC}.$$

However, the operator $\Omega_{AA'}$ only acts on λ^A if it differentiates in a direction within the self-dual 2-plane corresponding to λ^A. If we wish $\Omega_{AA'}$ to provide a genuine connection on the spin-bundle, we require its definition to be generalized so that it can act in *any* direction. We can take

$$\Omega_{AA'} = \nabla_{AA'} - \lambda^C\,\rho_{A'AC}.$$

Note that in the above definition of $\Omega_{AA'}$ it is necessary that the part of $\rho_{A'AC}$ that is *skew* in AC be known, whereas for $\mathbf{W}_{A'}$ it was not. This comes about because we are now concerned with obtaining a genuine connection on the spin-bundle (albeit still a non-linear one) for which differentiation in directions other than just those compatible with λ^A are required. This suggests that the "$\rho_{A'AB}$" that is being used here should *not* be thought of as the direct analogue of the second potential arising in the *Dirac* formulation of helicity $^3/_2$, but of some kind of analogous second potential for the *Rarita-Schwinger* formulation.

We cannot simply ignore this skew part of $\rho_{A'AC}$ because the *gauge* transformations

$$\lambda_A \mapsto \tilde{\lambda}_A = \chi\,\lambda_A$$

are accompanied by

$$\rho_{A'AB} \mapsto \tilde{\rho}_{A'AB} = \rho_{A'AB} + \nabla_{AA'}\xi_B$$

where ξ_A depends only on the space-time point and not on the fibre coordinate, but where χ depends on both. Consistency requires

$$\chi = (1 + \lambda_A\xi^A)^{-1},$$

which gives our spin-bundle a somewhat "crazy" non-linear structure.

7. Description in terms of a vector bundle of rank 3

Rather than having non-linear fibres, we might prefer to pass to a description in terms of a genuine vector bundle of rank 3, as follows. In place of a fibre coordinate λ_A (rank 2) we can use

$$(\eta_A, \zeta) \quad \text{where} \quad \lambda_A = \zeta^{-1}\eta_A.$$

Then

$$\Omega_{AA'}(\eta_B, \zeta) = (\nabla_{AA'}\eta_B, \ \nabla_{AA'}\zeta - \eta^C\rho_{A'AC}).$$

The gauge transformations are

$$(\eta_B, \zeta) \mapsto (\eta_B, \zeta) = (\eta_B, \ \zeta + \eta_A\xi^A).$$

Note that the gauge freedom on $\rho_{A'AB}$, namely

$$\rho_{A'AB} \mapsto \rho_{A'AB} + \nabla_{AA'}\xi_B,$$

gives

$$\rho_{A'(AB)} \mapsto \rho_{A'(AB)} + \nabla_{A(A'}\xi_{B)}$$

(which preserves the "field" $\varphi_{A'B'C'} = \nabla^{B}{}_{(B'}\nabla^{C}{}_{C'}\rho_{A')BC}$, in flat space-time) and

$$\rho_{A'C}{}^{C} \mapsto \rho_{A'C}{}^{C} + \nabla_{CA'}\xi^{C}.$$

The condition for $\Omega_{AA'}$ to be integrable on β-planes is

$$\nabla^{A'(A}\rho_{A'}{}^{B)C} = 0.$$

Note that, on the symmetric and skew parts of $\rho_{A'}{}^{BC}$ this is

$$\nabla^{A'(A}\rho_{A'}{}^{BC)} = 0$$

and

$$\nabla^{A'A}\rho_{A'(AB)} = {}^{3}/_{2}\ \nabla^{A'}{}_{B}\,\rho_{A'C}{}^{C}.$$

Note that $\rho_{A'C}{}^{C}$ constitutes an additional "gauge", which does not contribute to the field of helicity $^{3}/_{2}$.

This appears to be much more within the framework of the Rarita-Schwinger, rather than the Dirac formulation, and we may ask what should be the corresponding Rarita-Schwinger *first* potential $\sigma^{A}{}_{B'C'}$ corresponding to this $\rho^{BC}{}_{A'}$. In fact we find, rather surprisingly, that if $\sigma^{A}{}_{B'C'}$ is defined according to

$$\sigma^{A}{}_{B'C'} = \nabla_{CC'}\,\rho_{B'}{}^{AC}$$

(note ordering of indices!), then the Rarita-Schwinger equations

$$\nabla^{B'(C}\sigma^{A)}{}_{B'C'} = 0 \text{ and } \nabla_{C(C'}\sigma^{C}{}_{A')}{}^{C'} = 0$$

are satisfied in flat (or at least self-dual) space-time.

8. Further comments

Since the lecture on which this article is based was given, certain advances have been made. A long exact sequence has been found for the first and second Rarita-Schwinger-type potentials $\sigma^{A}{}_{B'C'}$ and $\rho_{A'}{}^{BC}$ given here. This has also been generalized, using a local-twistor formulation due to Lionel Mason, to give a more general form for $\rho_{A'}{}^{BC}$ in which its gauge freedom is unrestricted. (See Mason and Penrose 1994 for further details.) The active role of this extended Rarita-Schwinger-type second potential, and some of its interrelations with general relativity, are being explored.

References

Buchdahl, H.A. (1958) On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field *Nuovo Cim.* **10**, 96-103.

Deser, S. and Zumino, B. (1976) Consistent supergravity *Phys. Lett.* **62B**, 335-7.

Dirac, P.A.M. (1936) Relativistic wave equations *Proc. Roy. Soc. (Lond.)* **A155**, 447-59.

Eastwood M.G., Penrose, R., and Wells, R.O., Jr. (1981) Cohomology and massless fields, *Comm. Math. Phys.* **78**, 305-51.

Julia, B. (1982) Système linéare associé aux équations d'Einstein Comptes. Rendus. Acad. Sci. Paris, Sér. II, 113-6.

LeBrun, C.R. (1990) Twistors, ambitwistors, and conformal gravity, in *Twistors in Mathematical Physics,* LMS Lec. Notes **156**, eds T.N. Bailey and R.J. Baston (Cambridge Univ. Press, Cambridge).

Mason, L.J. (1985) The structure and evolution of hypersurface twistor spaces. *Twistor Newsletter* **20**, 75-79.

Mason, L.J. and Penrose, R. (1989) A twistorial approach to the full vacuum equations. *Twistor Newsletter* **29**, 1-5.

Mason, L.J. and Penrose, R. (1994) Spin 3/2 fields and local twistors. Twistor Newsletter **37**, 1-6.

Newman, E.T. (1976) Heaven and its properties *Gen. Rel. Grav.* **7**, 107-11.

Penrose, R. (1969) Solutions of the zero rest-mass equations, J. *Math. Phys.* **10**, 38-9.

Penrose, R. (1976) Non-Linear gravitons and curved twistor theory *Gen. Rel. Grav.* **7**, 31-52.

Penrose, R. (1986) On the origins of twistor theory, in *Gravitation and Geometry,* (I. Robinson festschrift volume), eds. W. Rindler and A. Trautman (Bibliopolis, Naples).

Penrose, R. (1990a) Twistor theory after 25 years — its physical status and prospects, in *Twistors in Mathematical Physics,* LMS Lec. note ser. 156, eds T.N. Bailey and R.J. Baston (Cambridge Univ. Press, Cambridge).

Penrose, R (1990b) Twistors theory for vacuum space-times: a new approach. *Twistor Newsletter* **31**, 6-8.

Penrose, R (1991) Twistors as spin 3/2 charges continued: SL(3,C) bundles. *Twistor Newsletter* **33**, 1-6.

Penrose, R.(1992a) $\mathcal{H}$-space and Twistors, in *Recent Advances in General Relativity,* (Einstein Studies, Vol. 4) eds. Allen I. Janis and John R. Porter (Birkhäuser, Boston) 6-25.

Penrose, R. (1992b) Twistors as spin 3/2 charges, in *Gravitation and Modern Cosmology* (P.G. Bergmann's 75th birthday vol.) eds. A. Zichichi, N. de Sabbata, and N.Sánchez (Plenum Press, New York).

Penrose, R. (1993) Quantum Non-Locality and Complex Reality, in *The Renaissance of General Relativity* (in honour of D.W. Sciama) eds. G. Ellis, A. Lanza, and J. Miller (Cambridge Univ. Press, Cambridge).

Penrose, R. and MacCallum, M.A.H. (1972) Twistor theory: an approach to the quantization of fields and space-time, *Phys. Repts.* **6C**, 241-315.

Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time,* Vol. 1: *Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge).

Penrose, R. and Rindler, W. (1986) *Spinors and Space-Time,* Vol. 2: *Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge.

Rarita, W and Schwinger, J. (1941) On the theory of particles with half-integer spin *Phys. Rev.* **60**, 61- .

Ward, R.S. (1979) The twisted photon: massless fields as bundles, in *Advances in Twistor Theory,* eds. L.P. Hughston and R.S. Ward (Pitman, London).

Ward, R.S. (1983) Stationary and axi-symmetric spacetimes, *Gen. Rel. Grav.* **15**, 105-9.

Woodhouse, N.M.J. and Mason, L.J. (1988) The Geroch group and non-Hausdorff twistor spaces *Nonlinearity* **1**, 73-114.

236
WHAT 'GAPS'? REPLY TO GRUSH AND CHURCHLAND

J. Consc. Stud. **2** (1995), 98–111. With S. Hameroff

As with Chapters 196, 206, 209, and 225, these responses appear without the critical remarks that engendered them. The citation for these, given in the reference list is: Grush, R. and Churchland, P. S. (1995) Gaps in Penrose's toilings. *J. Consc. Stud.* 2, 10–29.

Journal of Consciousness Studies, **2**, No.2, 1995, pp. 98–111

WHAT 'GAPS'? REPLY TO GRUSH AND CHURCHLAND

*Roger Penrose, Oxford University, Mathematical Institute,
24–29 St. Giles, Oxford OX1 3LB, UK
Stuart Hameroff, Departments of Anesthesiology and Psychology,
University of Arizona, Tucson, AZ 85724, USA. Email: srh@ccit.arizona.edu*

Abstract: Grush and Churchland (1995) attempt to address aspects of the proposal that we have been making concerning a possible physical mechanism underlying the phenomenon of consciousness. Unfortunately, they employ arguments that are highly misleading and, in some important respects, factually incorrect. Their article 'Gaps in Penrose's Toilings' is addressed specifically at the writings of one of us (Penrose), but since the particular model they attack is one put forward by both of us (Hameroff and Penrose, 1995; 1996), it is appropriate that we both reply; but since our individual remarks refer to different aspects of their criticism we are commenting on their article separately. The logical arguments discussed by Grush and Churchland, and the related physics are answered in Part 1 by Penrose, largely by pointing out precisely where these arguments have already been treated in detail in *Shadows of the Mind* (Penrose, 1994). In Part 2, Hameroff replies to various points on the biological side, showing for example how they have seriously misunderstood what they refer to as 'physiological evidence' regarding effects of the drug colchicine. The reply serves also to discuss aspects of our model 'orchestrated objective reduction in brain microtubules — **Orch OR**' which attempts to deal with the serious problems of consciousness more directly and completely than any previous theory.

Part 1: The Relevance of Logic and Physics

Logical arguments

It has been argued in the books by one of us, *The Emperor's New Mind* (Penrose, 1989 — henceforth *Emperor*) and *Shadows of the Mind* (Penrose, 1994 — henceforth *Shadows*) that Gödel's theorem shows that there must be something non-computational involved in mathematical thinking. The Grush and Churchland (1995 — henceforth G&C) discussion attempts to dismiss this argument from Gödel's theorem on certain grounds. However, the main points that they put forward are ones which have been amply addressed in *Shadows*. It is very hard to understand how G&C can make the claims that they do without giving any indication that virtually all their points are explicitly taken into account in *Shadows*. It might be the case that the arguments given in *Shadows* are in some respects inadequate, and it would have been interesting if G&C had provided a detailed commentary on these particular arguments, pointing out possible shortcomings where they occur. But it would seem from what G&C say that they have not even read, and certainly not understood, these arguments. A natural reaction to their commentary would be simply to say 'go and read the book and come back when you have understood its arguments.'[1] However, it will be helpful to pinpoint the specific issues that they raise here, and to point out the places in *Shadows* where these issues are addressed.

The authors are grateful to Vanessa Thomas for advice regarding the manuscript.

---

[1] It is extremely frustrating, considering the efforts involved in writing a book with particularly detailed arguments, when these arguments are simply treated as though they did not exist! Grush and Churchland are by no means unique in their ignoring most of the detailed arguments given in *Shadows*. For example, in their footnote 9, page 13, they refer to the 'powerful criticism of Putnam' (1994), who equally seems not even to have read the arguments in *Shadows* of relevance to his discussion.

The main argument that they appear to be raising against Penrose's (1989; 1994) use of Gödel's theorem (to demonstrate non-computability in mathematical thinking) is that mathematical thinking contains errors. They give the impression that the possibility of errors by mathematicians is not even considered by Penrose. However, in §§3.2, 3.4, 3.17, 3.19, 3.20 and 3.21 of *Shadows* the question of possible errors in human or robot mathematical reasoning is explicitly addressed at length. (The words 'errors' and 'erroneous' even appear explicitly in the headings of two of those sections and it is hard to see why G&C make no reference to these parts of the book.) In addition, on page 16 of their commentary, G&C claim that 'most of the technical machinery' involved in Penrose's arguments refer to what they call 'A1a' and 'A1c', on their page 15, which they choose not to dispute; whereas in fact by far the most difficult technical arguments given in *Shadows* are those which *specifically* address the possibility of *errors* in human or robot mathematical reasoning (these are given in §§ 3.19 and 3.20 of *Shadows*). It is difficult to understand why G&C fail to refer to this discussion, seeming to suggest (quite incorrectly) that Penrose has an in-built faith in the complete accuracy in the reasoning of mathematicians! G&C have a curious way of formalizing what they believe to be the ingredients of Penrose's arguments. In particular, on page 16 they refer to 'Penrose's Premise A1: Human thought, at least in some instances, *perhaps in all*, is sound, yet non-algorithmic' (which they break down into A1a, . . . , A1e). Their 'Premise A1' is nowhere to be found in Penrose's writings. It is fully admitted by Penrose that actual human thinking can be unsound even when seeming to be carried out in the most rigorous fashion by mathematicians.

It may well be that there is a genuine and deep misunderstanding implicit in what G&C are attempting to say, and it may be helpful to try to clarify the issue here. For the purposes of our present discussion (and for the essential discussion given in *Shadows*) it will be sufficient to restrict attention to a very specific class of mathematical statements, namely those referred to as Π_1-sentences. Such sentences are assertions that particular (Turing-machine) computations do not halt.

There are some very famous examples of mathematical assertions which take the form of Π_1-sentences, the best known being the so-called 'Fermat's Last Theorem'. Other examples are 'Goldbach's conjecture' (still unproved) that every even number greater than 2 is the sum of two primes, Lagrange's Theorem that every natural number is the sum of four squares, and the famous 4-colour theorem. It is useful to concentrate one's attention on Π_1-sentences because this is all one needs for application of the Gödel argument to the issue of computability in human mathematical thinking. There is no relevant issue of dispute between mathematicians as to the meaningfulness and objectivity of the truth of such sentences. (One might, however, worry about the 'intuitionists' or other constructivists in this context — and some reference to such viewpoints is given on p. 18 and footnote 30 on p. 20 of the G&C article. However, such constructivist viewpoints do not evade the Gödel argument and the use made of it in *Shadows* as is explicitly addressed in the discussion of Q9 on page 87 of *Shadows*, a discussion not even referred to by G&C.)

As far as we can make out, G&C are not disputing the absolute ('Platonic') nature of the truth or falsity of explicit Π_1-sentences. The issue is the *accessibility* of the truth of Π_1-sentences by human reasoning and insight.

We should make clear what is meant by a word such as 'accessibility' in this context, since there seem to be a great many misconceptions by philosophers and others as to how mathematical understanding actually operates. It is *not* a question of some kind of 'mystical intuition' that (some) mathematicians might have, and which is unavailable to ordinary mortals. What is being referred to by 'access' is simply the normal procedure of *mathematical proof.* It is not even a question of how some mathematician might have the inspiration to arrive at a proof. It is merely the question of the understanding which is involved in the ability to follow a proof in principle. (See, in particular, in the response to Q12 pp. 101–3 of *Shadows*.) However, it should be made clear in this context that the word 'proof' does not refer necessarily to a formalized argument within some preassigned logical scheme. For example, the arguments given by Andrew Wiles (as completed by Taylor and Wiles) to demonstrate the validity of Fermat's last theorem were certainly not presented as formal arguments, within, say, the Zermelo-Fraenkel axiom system. The essential point about such arguments is that they have to be *correct as mathematical reasoning*. It is a secondary matter to try to find out within which formal mathematical systems such arguments can be formulated. Indeed, what the Gödel argument shows (and this is not in dispute) is that if the rules of some formal system, **F**, can be trusted as providing correct demonstrations of mathematical statements — and here we need restrict attention only to Π_1-sentences — then the particular Π_1-sentence G(**F**) must also be accepted as true even though it is not a consequence of the very rules provided by **F**. (Here the sentence G(**F**) is the Gödel proposition which asserts the consistency of the formal system **F** — assuming that **F** is sufficiently extensive. It can also be taken as the explicit statement $C_k(k)$ exhibited on p. 75 of *Shadows*.[2] What this shows is that mathematical understanding (i.e. mathematical proof — in the sense above) cannot be encapsulated in any humanly acceptable formal system. Here 'acceptable' means acceptable to mathematicians as a reliable means of obtaining mathematical truths, where attention may be restricted to the truth of Π_1-sentences.

The notion of 'proof' that is being referred to above certainly raises profound issues. However, it would be unreasonable to dismiss it as something which is too ill-defined for scientific consideration or perhaps 'mystical'. There is indeed something mysterious about the very nature of 'understanding' and this is what is involved here. But the notion of proof that is involved in mathematical understanding is extraordinarily precise and accurate. There is no other form of argument within science or philosophy which really bears comparison with it. Moreover, this notion transcends any individual mathematician. But it is what mathematicians individually strive for. If one mathematician claims to have an argument for demonstrating the validity of some assertion — say a Π_1-sentence — then it should in principle be possible to convince another mathematician that the argument, and hence the conclusion, is correct — unless there is an error, in which case it is up to the mathematicians to locate this error. There is no question but that mathematicians do, not infrequently, make errors. This is not the point. The point is that it is possible for there actually to *be* an argument of some nature, to be found, which demonstrates the truth of the Π_1-sentence in question, and it is the mathematicians' business to try to find such an argument, whether or not they make mistakes in the process of attempting to achieve this. If there is an argument accessible to human understanding, then 'access' to this particular Π_1-sentence is possible. The point about the Gödel argument in this context is that there is no way to encapsulate the means whereby this access is achieved, in terms of mathematically acceptable computational rules.

---

[2] There is an unfortunate error in *Shadows* which arises from a misunderstanding on the part of the author about the precise relationship between the specific statement that Gödel originally produced and the condition of Ω-consistency for a formal system, for which the notation 'Ω(**F**)' is used. This has been corrected in the paperback edition, but the alteration makes virtually no change to the arguments. The simplest way for a reader in possession of the original hardback printing to deal with the point is simply to replace each occurrence of Ω(**F**) in Chapter 3 with G(**F**).

It is no part of Penrose's argument against computationalism that *all* Π_1-sentences should be humanly accessible (although there is some discussion of this possibility in chapter 8 of *Shadows*). What is argued for in *Shadows* is the assertion that the class of Π_1-sentences which are in principle humanly accessible is *not* a class which is computationally accessible (technically; *not a recursive set* — it is certainly not a knowably recursive set).

The issue of *practical* accessibility (as opposed to *in principle* accessibility) of Π_1-sentences by individual mathematicians, or perhaps by the mathematical community as a whole, is of course a somewhat separate matter, but these issues are discussed in considerable detail in Chapter 3 of *Shadows* (and also in the discussion of Q8 on pp. 83–9). No mention of this is made by G&C. Nor do they refer to the fact that in *Shadows* the Gödel argument is applied to computability *both* in the 'in principle' and the 'in practice' sense, according to context.

A detailed commentary on all of the remaining misunderstandings of Penrose's arguments displayed by G&C would take more space than is available to us here. However, it will be helpful to list some of the more important of these. At the end of the paragraph which finishes at the top of page 17, in a rather confused sentence, they seem to be saying that there are no sound procedures of mathematical deliberation, beyond, say, Zermelo-Fraenkel set theory (**ZF**). This shows that they do not understand Gödel. He demonstrated that if **ZF** is sound, then G(**ZF**) does indeed enable one to go beyond **ZF** (etc.). On page 18 they refer to the changing perceptions of mathematical rigour over the centuries, seeming to suggest that Cauchy and Cantor might have had different conceptions of what is unassailably true in mathematics. (Here, 'unassailable' refers merely to 'correctly proved' in the sense given above.) There is nothing wrong with the continual broadening of the kind of reasoning which can be used to obtain mathematical results, say Π_1-sentences, and what Cantor did was to increase this breadth enormously over what had been achieved before. This led to delicate, and sometimes disputed issues, such as the Axiom of Choice. The possibility of differing viewpoints with regard to the Axiom of Choice is explicitly addressed in the discussion of Q11 on pp. 97–101 of *Shadows* (although G&C seem oblivious of this fact). The conclusion is that this does not significantly affect the non-computability argument.

G&C make a number of points which seem to suggest that they think that Penrose is unaware of certain elementary logical points. This is particularly irritating. For example at the bottom of p. 19 they point out that: 'M does *not* believe that A is unassailably true' does not entail that 'M *dis*believes A'. Of course it does not. This kind of distinction is essential to the arguments of Chapter 3 of *Shadows*.

Physical arguments

There is a basic confusion in G&C p. 15, concerning the algorithmic or non-algorithmic nature of physical laws. The issue of approximations is brought out in *Shadows* only as a stop-gap measure to handle the *continuous* parameters in terms of which modern physical theories are invariably described. Artificial neural networks (ANNs) *are* simply computational devices, being digitally, computationally controlled, and there is no issue of approximation involved. Moreover, there is surely no serious suggestion that the present-day ANNs are actually conscious, or possess genuine understanding, despite all their 'successes'. The Penrose argument implies that we must go beyond such computational action if we are to find a physical basis for consciousness. ANNs are extensively considered in *Shadows*, and constitute an important part of the arguments of Chapter 3.

G&C spend some time in their article attacking the suggestion that the growth of quasicrystals might be non-computational. This is a red herring. It should be made clear that such a suggestion is nowhere made in Penrose's writings. In *Emperor* (but not in *Shadows*) quasicrystals are mentioned, and the suggestion is made that there might be something essentially *non-local* (not non-computational) in quasicrystal assembly which could perhaps involve the unknown physics underlying quantum state reduction in an essential way.

G&C claim that there is no physical indication that the unknown theory of 'quantum gravity' (which correctly combines gravitational theory with quantum mechanics) should be non-computational. Penrose does not make any **strong** claim for such non-computability from purely physical theory. However, despite what G&C say, such evidence *does* exist and is discussed in §§7.8 and 7.10 in *Shadows*.

Finally, we come to what G&C call the 'convenient myth' of Platonism (p. 21), where they refer to Fig 8.1 on p. 414 of *Shadows*. The main worry that they (and others) seem to have about Platonism relates to 'mystery 3' of that diagram, namely the relationship between human thought and absolute mathematical truth.

However, the real mystery for mathematical physicists is 'mystery 1': how is it that the physical world indeed accords — and has accorded since the beginning of time — with such extraordinary precision with subtle and beautiful mathematical laws. If these mathematical laws are merely the product of our recent mental activity, then we are presented with a profound paradox.

Part 2: The Biological Side

Remarks by G&C pertaining to microtubules (MTs), quantum theory and consciousness fall into the following areas:

1. How could quantum gravity, MTs and consciousness 'conceivably' be inter-related?
2. How could consciousness depend on MTs when 'physiological evidence demonstrates' that consciousness can occur without them?
3. How do MTs communicate with neural membrane and synaptic functions? How can MTs encode and process information?
4. Given the apparently noisy, thermal environment within neurons and the brain, how could quantum coherent phenomena occur a) within neurons, and b) throughout macroscopic brain regions?
5. Why don't ions such as sodium and calcium prevent quantum phenomena in cytoplasm? Within hollow MT cores?
6. Miscellaneous

1. How could quantum gravity, MTs and consciousness 'conceivably' be inter-related?

Since papers describing our model of 'orchestrated objective reduction in brain microtubules — **Orch OR**' (Hameroff and Penrose, 1995; 1996; forthcoming) had not been published, nor reviewed by G&C at the time their critique was written, we summarize key points of the model here:

(1) Aspects of quantum theory (e.g. quantum coherence) and of the suggested physical phenomenon of quantum wave function 'self-collapse' (objective reduction: **OR** — Penrose, 1994) are essential for consciousness, and occur in cytoskeletal microtubules (MTs) and other structures within each of the brain's neurons.

(2) Conformational states of MT subunits (tubulins) are coupled to internal quantum events, and cooperatively interact with other tubulins in both classical and quantum computation.

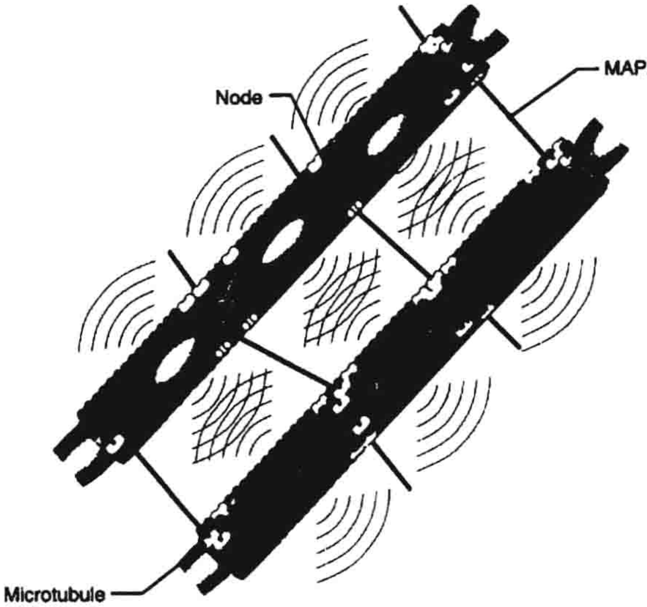


Figure 1. Schematic representation of proposed quantum coherence in microtubules. Having emerged from resonance in classical automaton patterns, quantum coherence non-locally links superpositioned tubulins (grey) within and among microtubules. Upper microtubule: cutaway view shows coherent photons generated by quantum ordering of water on tubulin surfaces, propagating in microtubule waveguide. MAP (microtubule-associated-protein) attachments breach isolation and prevent quantum coherence; MAP attachment sites thus act as 'nodes' which tune and orchestrate quantum oscillations and set possibilities and probabilities for collapse outcomes ('orchestrated objective reduction — **Orch OR**').

(3) Quantum coherence occurs among tubulins in MTs, pumped by thermal and biochemical energies (e.g. Fröhlich, 1968; 1970; 1975). Water at MT surfaces is 'ordered' — dynamically coupled to the protein surface. Water ordering within the hollow MT core (acting like a quantum waveguide) can result in quantum coherent photons ('super-radiance' and 'self-induced transparency' — Jibu *et al.*, 1994; 1995).

(4) During pre-conscious processing, quantum coherent superposition/computation occurs in MT tubulins and continues until the mass-distribution difference among the separated states of tubulins reaches a threshold related to quantum gravity. Self-collapse (**OR**) then occurs.

(5) The **OR** self-collapse process results in classical 'outcome states' of MT tubulins which then implement neurophysiological functions. According to principles of **OR** (Penrose, 1994), the outcome states are 'non-computable' — they cannot be determined algorithmically from the tubulin states at the beginning of the quantum computation.

(6) Possibilities and probabilities for post-**OR** tubulin states are influenced by factors including initial tubulin states, and attachments of microtubule-associated proteins (MAPs) acting as 'nodes' which tune and 'orchestrate' the quantum oscillations (Fig. 1). We thus term this self-tuning **OR** process 'orchestrated objective reduction — **Orch OR**'.

(7) According to the arguments for **OR** put forth in Penrose (1994), superpositioned states each have their own space-time geometries. When the degree of coherent mass-energy difference leads to sufficient separation of space-time geometry, the system must choose and decay (reduce, collapse) to a single universe state. Thus **Orch OR** involves self-selections in fundamental space-time geometry (Hameroff and Penrose, forthcoming).

(8) To quantify the **Orch OR** process, we calculate from the indeterminacy principle $E=\hbar/T$, in which E is the gravitational self-energy of the quantum superposition (a mass acting on its displaced self), $\hbar$ is Planck's constant over 2π, and T is the coherence time (how long the coherence is sustained). If we assume a coherence time T = 500 msec (shown by Libet *et al.*, 1979, and others to be a relevant time for pre-conscious processing), we calculate E, and determine the number of MT tubulins whose coherent superposition for 500 msec will elicit **Orch OR**. This turns out to be about 10^9 tubulins.

(9) A typical brain neuron has roughly 10^7 tubulins (Yu and Baas, 1994). If, say, 10 percent of tubulins within each neuron are involved in the quantum coherent state, then roughly 10^3 (one thousand) neurons would be required to sustain coherence for 500 msec, at which time the quantum gravity threshold is reached and **Orch OR** then occurs.

(10) We consider each self-organized **Orch OR** as a single conscious event; a series of such events would constitute a 'stream' of consciousness. If we assume some form of excitatory input (e.g. you are threatened, or enchanted) in which quantum coherence emerges faster, then, for example, 10^{10} coherent tubulins could **Orch OR** after 50 msec, or 10^{11} after 5 msec. Turning to see a Bengal tiger in your face might perhaps elicit 10^{12} in 0.5 msec, or more tubulins, faster.[3] A slow emergence of coherence (your forgotten phone bill) may require longer times.

(11) Each instantaneous **Orch OR** may then 'bind' various superpositions which may have evolved in separated spatial distributions and over different time scales, but whose net displacement self-energy reaches threshold at a particular moment. Information is bound into an instantaneous event (a 'conscious now'). Cascades of **Orch OR**s define a forward flow, or stream of consciousness.

The **Orch OR** model thus accommodates some important features of consciousness: 1) control/regulation of neural action, 2) pre-conscious to conscious transition, 3) non-computability, 4) causality, 5) binding of various (time scale and spatial) superpositions into instantaneous 'now', 6) a 'flow' of time, and 7) a connection to fundamental space-time geometry.

2. How could consciousness depend on MTs when 'physiological evidence demonstrates' that consciousness can occur without them?

The 'physiological evidence' provided by G&C turns out to be:

1. the drug colchicine causes depolymerization of MTs
2. patients taking colchicine pills for treatment of gout, and experimental animals receiving colchicine directly to the brain or cerebrospinal fluid don't become unconscious. (G&C fail to mention that the animals *do* become 'demented'.)

The drug colchicine (from the plant *Colchicum automnale*) has been known as a poison since antiquity, and used as medicine since the 18th century (Dustin, 1984). It is still used for the treatment of gout, first described by Hippocrates (fifth century BC) and caused by accumulation of urate crystals in joint spaces. Immune cells (lymphocytes and macrophages) migrate to, and engulf the 'foreign' crystals. The cells release inflammatory substances which cause the joint swelling and pain characteristic of gout. Cell migration and other activities depend on dynamic rearrangements of their cytoskeletons, including, and depending primarily on, MTs. For example in locomotion, MTs inside the cells self-assemble in the proper orientation; other cell components then follow or are transported.

Colchicine prevents the immune cell locomotion and other activities by binding to growing ('beta plus') ends of actively polymerizing MTs. This prevents the cycles of MT polymerization and depolymerizeration required during cell division (MT 'mitotic spindles'), and locomotion and other immune cell activities responsible for symptoms of gout and other inflammatory processes. In patients taking colchicine pills for gout, drug access to the brain through the blood-brain barrier is about 10^{-4} of blood levels (Bennett *et al.*, 1981). So it's not surprising in any case that these patients have no central nervous system symptoms.

---

[3] The faster emergence we describe is distinct from 'reflex' phenomena: touching your finger to a hot stove may result in spinally mediated withdrawal before conscious awareness of the pain.

G&C cite studies (e.g. Bensimon and Chernat, 1991; Kolasa *et al.*, 1992; Emerich and Walsh, 1991) in which colchicine delivered directly to the brain or cerebrospinal fluid of experimental animals failed to cause loss of consciousness. G&C neglect to mention that the colchicine does cause significant cognitive impairments of learning and memory. Bensimon and Chernat (1991), for example, characterized a 'dementia' in their colchicine-treated animals which they likened to Alzheimer's disease (often linked to MT disruption — e.g. Matsuyama and Jarvik, 1992).

G&C are thus asking: if colchicine depolymerizes MTs, why doesn't it cause unconsciousness (rather than just dementia)?

Answer: even when given in huge doses directly to the brain, colchicine has minimal depolymerizing effects on brain MTs.

Unlike the labile MTs in cell division and locomotion, most MTs in brain neurons are stable, 'hardened' by biochemical changes, do not engage in cycles of polymerization and depolymerization, and are resistant to colchicine. They have no exposed 'beta plus ends'!

Dynamic neuronal MTs which *do* undergo repeating cycles of polymerization and depolymerization are those involved in restructuring synaptic connections, perhaps accounting for colchicine's impairments of learning and memory.

3. How do MTs communicate with neural membrane and synaptic functions? How can MTs encode and process information?

This topic is discussed in great detail in Hameroff (1987), Rasmussen *et al.* (1990), Dayhoff *et al.* (1994) and Hameroff and Penrose (1996).

As a general summary, MTs are linked to membrane receptors and ion channels both structurally, via smaller cytoskeletal proteins like fodrin, actin, synapsin and others, and biochemically, for example as part of second messenger cascades. MT actions define and modify neural architecture and synaptic connections and strengths. The MT cytoskeleton is each neuron's nervous system.

As one example, consider 'MAP-2', a dendrite-specific, MT-crosslinking 'microtubule-associated-protein' ('MAP'), necessary for learning and memory. As a result of synaptic membrane receptor activation (e.g. Halpain and Greengard, 1990), MAP-2 is 'dephosphorylated' (imparting energy and information to the cytoskeleton). This process is essential to strengthening synaptic pathways, for example in cat visual cortex with visual stimulation (Aoki and Siekevitz, 1985) and rat temporal cortex in auditory Pavlovian conditioning (Woolf *et al.*, 1994). MAP-2 dephosphorylation, which consumes a large proportion of brain biochemical energy (e.g. Theurkauf and Vallee, 1983) acts to reconfigure the sub-synaptic cytoskeleton (Bigot and Hunt, 1990; Friedrich, 1990).

Regarding neurotransmitter release, it is true as G&C say that MTs don't extend into the terminal axon region (nor into dendritic spines — MTs are too large). MTs do, however, connect to smaller cytoskeletal proteins like synapsin which, influenced by calcium ion fluxes, are directly involved in neurotransmitter release (Hirokawa, 1991). (In dendrites, MTs connect to smaller actin filaments which extend into dendritic spines and interact with receptors.) As Beck and Eccles (1992) point out, the process of neurotransmitter release has a seemingly random, probabilistic component (only about one sixth of axonal depolarizations result in neurotransmitter vesicle release). Beck and Eccles suggest this may reflect some unrecognized quantum influence, although we don't share their view that it is a purely 'dualist' influence.

How can MTs encode, and process information? Vassilev *et al.* (1985) demonstrated signal transmission along tubulin chains. As cylindrical lattices of tubulin dipoles, MTs are well suited to process information, and a number of models of MT signalling and information processing have been suggested. These include propagating tubulin conformational changes (Atema, 1973; Roth and Pihlaja, 1977; Hameroff and Watt, 1982), sequential phosphorylation/dephosphorylation along MT tubulins (Puck and Krystosek, 1992), tensegrity (Wang and Ingber, 1994), non-linear soliton waves (Chou *et al.*, 1994; Sataric *et al.*, 1992), spin-glass and ferroelectric effects (Tuszynski *et al.*, 1995), 'cellular automaton' behavior (e.g. Rasmussen *et al.*, 1990) and quantum coherent photons (Jibu *et al.*, 1994; 1995).

4. Given the apparently noisy, thermal environment within neurons and the brain, how could quantum coherent phenomena occur: a) within neurons, and b) throughout macroscopic brain regions?

This is the crux of the matter. How can quantum effects emerge and be sustained in or around neurons for physiological time durations? The cellular milieu is considered noisy, thermal and chaotic. Ionic fluxes such as axonal depolarization would seem to disrupt quantum coherent effects, if they did occur. We consider 3 types of quantum phenomena:

Quantum coherent superposition ('Fröhlich mechanism') leading to self-collapse (**Orch OR**) among hydrophobic pockets in MT tubulins (Hameroff and Penrose, 1995; 1996).

Dynamic ordering of water on MT outer surfaces, in cytoplasm and extracellular spaces (Ricciardi and Umezawa, 1967; Del Giudice *et al.*, 1983).

Coherent photons ('super-radiance and self-induced transparency') in hollow MT 'waveguides' (Jibu *et al.*, 1994; 1995).

The internal cell environment in which MTs (and life) exist is the cytoplasm. The cytoplasm within neurons and the surrounding extracellular space differs markedly from non-living aqueous media, considered as water molecules in continuous thermal agitation. For example, proteins dissolved in the cytoplasm may be triggered by ionic fluxes to assemble rapidly into 'gelatinous' layers or structures. Cell cytoplasm may thus fluctuate between liquid solution ('sol'), and more solid-state gelatin ('gel').

Extensive charged surfaces on the cytoskeleton, membranes, organelles and extra-cellular matrices bind and order cytoplasmic water, which engage in cooperative dynamics. Several layers of ordered water on each of these many surfaces are predicted (e.g. Clegg, 1983), so that large proportions of cell interiors may be either dynamically ordered, gelatinous or solid.

According to Fröhlich (1968; 1970; 1975), dipole biomolecules structurally confined in membranes, MTs (and ordered water on their surfaces) become excited coherently by biochemical and thermal energy. The excitations reduce to a common frequency mode (10^9 to 10^{11} Hz), somewhat like the quantum phenomenon of a Bose-Einstein condensate (Anderson *et al.*, 1995). In Bose-Einstein condensates like superconductors, coherence is attained by extreme cooling to remove thermal vibrations; in lasers, and in the Fröhich model, the coherence derives from energy pumping.

Fröhlich coherence among (hydrophobic pockets within) MT subunits has been proposed as a basis for information processing via neighbor tubulin dipole interactions (e.g. as in a 'cellular automaton': Rasmussen *et al.*, 1990). The coherent dynamics are viewed also to order water at MT outer and inner surfaces; the cytoplasm can transiently assume a quantum coherent state.

Jibu *et al.* (1995) have recently examined Albrecht-Buehler's (1992) experimental results in which single cells detect, orient and move toward weak red/infra-red light signals using their cytoskeleton. They explain the surpisingly efficient photon propagation through cytoplasm and extra-cellular fluid by dynamical water ordering within, and outside the cell. According to Jibu *et al.*, the ordering behaves as a nonlinear coherent optical device — a 'water laser' — which enables lossless propagation of quantum coherent photons for a distance of 50 microns. Larger macroscopic quantum states may occur by coalescence of these regions, or by means such as quantum electromagnetic effects (Jibu *et al.*, 1995).

The Jibu *et al.* calculation of 50 microns implies quantum coherent regions with volumes of roughly 10^{-6} ml, or 10^{-9} litre. If one estimates brain volume of one litre, this fraction (10^{-9}) of total brain is about the same as our calculated 10^9 (out of 10^{18} to 10^{19} total) brain tubulins, and could represent a fundamental unit of consciousness: the gravitational self-energy E, for a given coherence time (in this case, T = 500 msec). E, of course, will vary inversely with T, akin to the relationship between wavelength and frequency in the electromagnetic spectrum.

In summary, MTs and associated water can shield or isolate quantum coherence/self collapse (**Orch OR**) from thermal disruption by several possible means:

1. Water in cells, or at least several water layers at MT and other surfaces, becomes dynamically ordered, and participates in quantum states.
2. Hydrophobic pockets within each tubulin are shielded from water, and thermal energy condenses to a coherent state akin to a Bose- Einstein condensate, as Fröhlich (1968; 1970; 1975) has suggested. Quantum events within the hydrophobic pockets can influence protein conformation, and MT function.
3. Calcium coupled sol-gel transformations form 'gelatinous' layers which transiently isolate MTs (Hameroff and Penrose, 1996).
4. The hollow MT inner core offers several theoretical advantages: it is sheltered, the particular quantum phenomena predicted to occur there (super-radiance and self-induced transparency) are quicker than thermal effects, and the cylinder can act as a quantum optical waveguide.

5. *Why don't ions such as sodium and calcium prevent quantum phenomena in cytoplasm? Within hollow MT cores?*

Effects of ions on the dynamically ordered structure of water depend on the relative size of each ion compared to the water molecule H_2O (Ergin, 1983; Uedaira and Osaka, 1989; Jibu *et al.*, 1995;). Ions whose radii are smaller than the H_2O radius (1.38 angstroms, or Å) do not disturb the ordering. Sodium ion (radius 0.98Å), calcium (1.00Å) and magnesium (0.72Å) can all embed between ordered H_2O molecules without disturbing them. Ions whose radii are close to that of water can take part in the dynamical geometry by replacing a water molecule. Potassium (1.33Å) is in this category. Ions larger than H_2O will disturb ordering. Chloride (1.81Å) should therefore disrupt 'Type 2' phenomena. (Intra-neuronal chloride concentration is extremely low, however, except during the terminal phase of an action potential.)

Within the hollow MT core, the 'super-radiance and self-induced transparency' predicted in Jibu *et al.* (1994) have characteristic times (to complete one cycle of photon emission and transfer) much shorter than that of any thermally disordering phenomena.

In axons, chloride fluxes at the end of each action potential may 'decohere' Type 2 phenomena. In dendrites, graded potentials would seem to have less significant effects. We conclude that the quantum coherent superposition essential for consciousness occurs primarily in dendritic MTs. Recognition of dendrites as key regions related to consciousness is consistent with extensive work of Pribram (1991), Eccles (1992) and others.

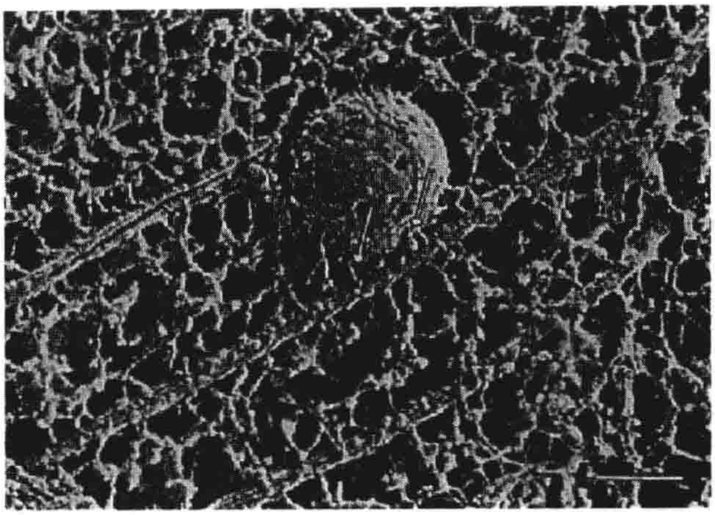


Figure 2.
Electron micrograph of axoplasmic transport of spherical vesicle by MTs. Approximately 3 MTs are visible, interconnected by MT-associated proteins ('MAPs'). The MTs are arrayed in parallel from lower left to upper right. Arrows indicate short cross-bridges between vesicle and MT. Scale bar (lower right) is 100 nanometers. Reprinted with permission from Hirokawa, 1991.

6. *Miscellaneous*

G&C p. 23: 'First we note that all cells have microtubules . . . ' [the implication presumably being 'why isn't your earlobe or some other bodily part conscious?'].

Penrose and Hameroff (P&H): Microtubules in neurons are quite distinct from those in other cells. 1) They are arrayed in parallel (rather than radially) because, unlike other cells, neurons lack centrioles. 2) They are quite stable. 3) They are far more abundant in neurons than other cells, form larger and complex networks, and have a greater genetic variability than other tissues (e.g. 17 isozymes of brain tubulin — Lee *et al.*, 1986). 4) They perform specialized transport along axon and dendrite processes. (Figure 2) 5) They have neuron-specific MT- associated proteins (MAPs).

G&C p. 23: The 'anaesthesia–microtubule connection.'

P&H: G&C imply that our view of the mechanism for general anaesthesia is depolymerization of MTs. This is incorrect. That idea emanated from Allison and Nunn (1968), in which the anaesthetic gas halothane depolymerized axopodia comprised of MTs in *Actinosphaerium* (but at 5 times the concentration required for clinical anaesthesia). Subsequent studies showed that at clinically relevant anaesthetic concentrations, MTs in neurons remain polymerized.

The mechanism of general anaesthesia is, we believe, an important clue to consciousness (e.g. Hameroff and Watt, 1983; Louria and Hameroff, 1995; Louria and Hameroff, 1996). Our view is that anaesthesia prevents quantum coherent superposition in hydrophobic pockets of a variety of neural proteins (receptors, ion channels, second messengers, enzymes, cytoskeletal MT subunits). Wulf and Featherstone (1967) showed that anaesthetic binding in internal protein hydrophobic pockets altered water binding at the protein outer surface, showing how the three types of quantum phenomena can be related.

G&C p. 24: 'There is no evidence that quantum coherence involving super-radiance (or anything else for that matter) occurs in microtubules. At best, what Hameroff [and Penrose have] . . . done is to show that it might be possible. This should most definitely be distinguished from providing *evidence* that it is *actual.*'

P&H: Agreed. It is a model. Presumably, future technologies will either refute or verify it. To put this in perspective, we could parody G&C as follows: 'There is no evidence that consciousness derives solely from activities at the level of neuronal assemblies. At best what the classical connectionists/reductionists have done is to show that it might be possible. This should most definitely be distinguished from providing *evidence* that it is *actual.*'

G&C p. 27: 'Despite the rather breathtaking flimsiness of the consciousness–quantum connection, the idea has enjoyed a surprisingly warm reception, at least outside neuroscience. One cannot help groping about for some explanation for this odd fact.' [After they isolate it, maybe they'll find a vaccine against it. 'Neuroscience' can still be saved.] 'Some people who, intellectually, are materialists, nevertheless . . . have a negative "gut" reaction to the idea that neurons . . . are the source of subjectivity and the "me-ness of me".'

P&H: Damasio (1994) suggests these people's emotional sub-conscious, mediated from the brain through the autonomic nervous system in their 'gut', is trying to tell them something. Our model suggests the emotional sub-conscious may derive from 'Platonic' quantum computing (*Shadows*, p. 414). Perhaps the 'explanation' is that some people are capable of perceiving subconscious 'gut' feelings, and others are not!

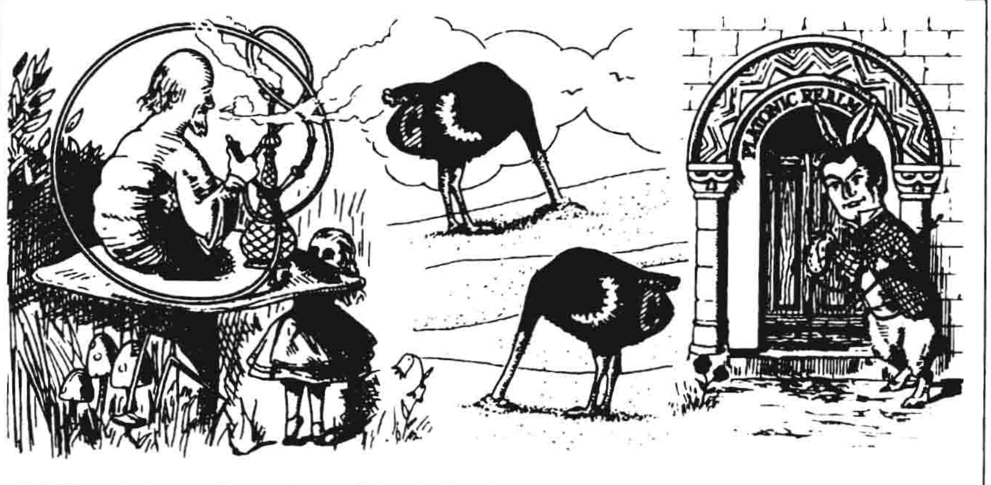
G&C p. 27: 'The crucial feature of neurons that makes them capable of processing and storing information is just ions passing back and forth across neuronal membranes through protein channels.'

P&H: Where is the evidence that neurons can process and store information without their cytoskeletons? More importantly, at this stage of the discussion, why can't G&C say that these ion/membrane activities account for consciousness? It must be because, as they say earlier (p. 10), 'Neuroscience has not reached the stage where we can satisfactorily answer these questions'. But even if every neuron, synapse, ion, channel and gene were mapped, would it necessarily tell us any more? Can the classical reductionists say whether *C elegans* (a nematode worm whose entire 302 neuron nervous system is completely mapped) is, or is not, conscious?

G&C p. 28: 'Why should it be less scary, reductionist or counter-intuitive that "me-ness" emerges from the collapse of a wave function than from neuronal activity?'

P&H: Microtubule functions *are* neuronal activities. However, it *is* somewhat appealing to see how the phenomenon of consciousness could tie in with the behaviour of the universe at its deepest levels, and be relevant even to the very geometrical structure of space-time.

G&C end by referring to Lewis Carroll's *Alice in Wonderland*, claiming our model (at that time not yet published, nor reviewed by G&C): 'is no better supported than any one of a gazillion caterpillar-with-hookah hypotheses.'



P&H: *It's not that we're in Wonderland*
But p'raps their heads are in the sand.

Conclusion

The Grush-Churchland arguments concerning the logical and physical parts of 'Penrose's viewpoint' do not accurately express this viewpoint, nor do they at all take into account thorough discussion of such arguments already given in *Shadows*. On the biological side they exhibit serious misunderstandings, for example in relation to what they deem 'physiological evidence' regarding effects of the drug colchicine. Nonetheless, we thank them for directing attention to the **Orch OR** model, which we believe deals with the serious problems of consciousness more directly and completely than any previous theory.

References

Albrecht-Buehler, G. (1992), 'Rudimentary form of cellular "vision" ', *Proc. Natl. Acad. Sci. USA*, **89** (17), pp. 8288–92.

Allison, A.C. and Nunn, J.F. (1968), 'Effects of general anesthetics on microtubules: a possible mechanism of anaesthesia', *Lancet*, **II**, pp. 1326–9.

Anderson, M.H., Ensher, J.R., Matthews, M.R., Wiseman, C.E. and Cornell, E.A. (1995), 'Observation of Bose-Einstein condensation in a dilute atomic vapor', *Science*, **269**, pp. 198–201.

Aoki, C. and Siekevitz, P. (1985), 'Ontogenic changes in the cyclic adenosine 3^1, 5^1 monophosphate — stimulatable phosphorylation of cat visual cortex proteins, particularly of microtubule-associated protein 2 (MAP2): effects of normal and dark rearing and of the exposure to light', *J. Neurosci.*, **5**, pp. 2465–83.

Atema, J. (1973), 'Microtubule theory of sensory transduction', *J. Theor. Biol.*, **38**, pp. 181–90.

Beck, F. and Eccles, J.C. (1992), 'Quantum aspects of brain activity and the role of consciousness', *Proc. Natl. Acad. Sci. USA*, **89** (23), pp. 11357–61.

Bennett, E., Alberti, M.H. and Flood, J., (1981), 'Uptake of [3H]Colchicine into brain and liver of mouse, rat, and chick', *Pharmacology, Biochemistry and Behavior*, **14** (6), pp. 863–9.

Bensimon, G. and Chernat, R. (1991), 'Microtubule disruption and cognitive defects: effect of colchicine on learning behavior in rats', *Pharmacol. Biochem. Behavior*, **38**, pp. 141–5.

Bigot, D. and Hunt, S.P. (1990), 'Effect of excitatory amino acids on microtubule-associated proteins in cultured cortical and spinal neurons', *Neurosci. Lett.*, **111**, pp. 275–80.

Chou, K-C., Zhang C-T. and Maggiora, G.M. (1994), 'Solitary wave dynamics as a mechanism for explaining the internal motion during microtubule growth', *Biopolymers*, **34**, pp. 143–53.

Clegg, J.S. (1983), 'Intracellular water, metabolism and cell architecture', in *Coherent Excitations in Biological Systems*, ed. H. Fröhlich and F. Kremer (Berlin: Springer-Verlag), pp. 162–75.

Damasio, A.R. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Putnam).

Dayhoff, J.E., Hameroff, S., Lahoz-Beltra, R. and Swenberg, C.E. (1994), 'Cytoskeletal involvement in neuronal learning: a review', *Eur. Biophys. J*, **23**, pp. 79–93.

Del Giudice, E., Doglia, S. and Milani, M. (1983), 'Self-focusing and ponderomotive forces of coherent electric waves: a mechanism for cytoskeleton formation and dynamics', in *Coherent Excitations in Biological Systems*, ed. H. Fröhlich and F. Kremer (Berlin: Springer-Verlag), pp. 123–7.

Dustin, P. (1984), *Microtubules* (2nd Revised Ed., Berlin: Springer).

Eccles, J.C. (1992), 'Evolution of consciousness', *Proc. Natl. Acad. Sci.*, **89**, pp. 7320–4.

Emerich, D.F. and Walsh, T.J. (1991), 'Ganglioside AGF2 prevents the cognitive impairments and cholinergic cell loss following intraventricular colchicine', *Experimental Neurology*, **112**, pp. 328–37.

Ergin, V. (1983), *Magintnye Svoistva i Struktura Rastvorov Elektritov* (Moscow: Nauka, in Russian).

Friedrich, P. (1990), 'Protein structure: the primary substrate for memory', *Neurosci.*, **35**, pp. 1–7.

Fröhlich, H. (1968), 'Long-range coherence and energy storage in biological systems', *Int. J. Quantum Chem.*, **2**, pp. 641–9.

Fröhlich, H. (1970), 'Long range coherence and the actions of enzymes', *Nature*, **228**, p. 1093.

Fröhlich, H. (1975), 'The extraordinary dielectric properties of biological materials and the action of enzymes', *Proc. Natl. Acad. Sci.*, **72**, pp. 4211–15.

Grush R., Churchland P.S. (1995), 'Gaps in Penrose's toilings', *J. Consciousness Studies*, **2** (1), pp. 10–29.

Halpain, S. and Greengard, P. (1990), 'Activation of NMDA receptors induces rapid dephosphorylation of the cytoskeletal protein MAP-2', *Neuron*, **5**, pp. 237–46.

Hameroff, S.R. and Watt, R.C. (1982), 'Information processing in microtubules', *J. Theor. Biol.*, **98**, pp. 549–61.

Hameroff, S.R. and Watt, R.C. (1983), 'Do anesthetics act by altering electron mobility?', *Anesth. Analg.*, **62**, pp. 936–40.

Hameroff, S.R. (1987), *Ultimate Computing: Biomolecular Consciousness and Nanotechnology* (Amsterdam: Elsevier North-Holland).

Hameroff S. and Penrose R. (1996), 'Orchestrated reduction of quantum coherence in brain microtubules — a model for consciousness', in *Toward a Science of Consciousness: Contributions from the 1994 Tucson Conference*, ed: S Hameroff, A Kaszniak and A Scott (Cambridge, MA: MIT Press).

Hameroff S. and Penrose, R. (1995), 'Orchestrated reduction of quantum coherence in brain microtubules', *Proc. Int. Neural Network Soc., Washington DC, July 17-21, 1995* (Hillsdale, N.J: Erlbaum).

Hameroff S. and Penrose, R. (forthcoming) 'Conscious events as orchestrated space-time selections', *Journal of Consciousness Studies*.

Hirokawa, N. (1991), 'Molecular architecture and dynamics of the neuronal cytoskeleton', in *The Neuronal Cytoskeleton*, ed. RD Burgoyne (New York: Wiley-Liss), pp. 5–74.

Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H. and Yasue, K. (1994), 'Quantum optical coherence in cytoskeletal microtubules: implications for brain function', *BioSystems*, **32**, pp. 195–209.

Jibu, M., Yasue, K. and Hagan, S. (1995), 'Water laser as cellular "vision"', Submitted.

Kolasa, K., Jope, R.S., Baird, M.S. and Johnson, G.V.W. (1992), 'Alterations of choline acetyltransferase, phosphoinositide hydrolysis, and cytoskeletal proteins in rat brain in response to colchicine administration', *Exp. Brain Res.*, 89, pp. 496–500.

Lee, J.C., Field, D.J., George, H.J. and Head, J. (1986), 'Biochemical and chemical properties of tubulin subspecies', *Ann. NY Acad. Sci.*, **466**, pp. 111–28.

Libet, B., Wright, E.W. Jr., Feinstein, B. and Pearl, D.K. (1979), 'Subjective referral of the timing for a conscious sensory experience', *Brain*, **102**, pp. 193–224.

Louria D. and Hameroff, S. (1995), 'Halothane-protein hydrophobic interactions: the quantum unitary hypothesis', *Anesthesiology*, in press.

Louria D. and Hameroff, S. (1996), 'Computer simulation of anesthetic binding in protein hydrophobic pockets', in *Toward a Science of Consciousness: Contributions from the 1994 Tucson Conference*, ed. S. Hameroff, A. Kaszniak and A. Scott (Cambridge, MA: MIT Press).

Matsuyama, S.S. and Jarvik, L.F. (1992), 'Hypothesis: Microtubules, a key to Alzheimer's disease', *Proc. Nat. Acad. Sci. USA*, **86**, pp. 8152–6.

Penrose, R. (1989), *The Emperor's New Mind* (Oxford: Oxford University Press).

Penrose, R. (1994), *Shadows of the Mind* (Oxford: Oxford University Press).

Pribram, K.H. (1991), *Brain and Perception* (Hillsdale, N.J: Lawrence Erlbaum).

Puck, T.T. and Krystosek, A. (1992), 'Role of the cytoskeleton in genome regulation and cancer', *Int. Rev. Cytology*, **132**, pp. 75–108.

Putnam, H. (1994), 'The best of all possible brains?', review of *Shadows of the Mind*, New York Times Review of Books, November.

Rasmussen, S., Karampurwala, H., Vaidyanath, R., Jensen, K.S. and Hameroff, S. (1990), 'Computational connectionism within neurons: a model of cytoskeletal automata subserving neural networks', *Physica D*, **42**, pp. 428–49.

Ricciardi, L.M. and Umezawa, H. (1967), 'Brain and physics of many-body problems', *Kybernetic*, **4**, 44–8.

Roth, L.E. and Pihlaja, D.J. (1977), 'Gradionation: hypothesis for positioning and patterning', *J. Protozoology*, **24** (1), pp. 2–9.

Sataric, M.V., Zakula, R.B. and Tuszynski, J.A. (1992), 'A model of the energy transfer mechanisms in microtubules involving a single soliton', *Nanobiology*, **1**, 445–56.

Theurkauf, W.E. and Vallee, R.B. (1983), 'Extensive cAMP-dependent and cAMP-independent phosphorylation of microtubule associated protein 2', *J. Biol. Chem.*, **258**, pp. 7883–6.

Tuszynski, J., Hameroff, S., Sataric, M.V., Trpisova, B. and Nip, M.L.A. (1995), 'Ferroelectric behavior in microtubule dipole lattices; implications for information processing, signaling and assembly/disassembly', *J. Theor. Biol.*, in press.

Uedaira, H. and Osaka, A. (1989), *Water in Biological Systems* (Tokyo: Kodansha Scientific, in Japanese).

Vassilev, P., Kanazirska, M. and Tien, H.T. (1985), 'Intermembrane linkage mediated by tubulin', *Biochem. Biophys. Res. Comm.*, **126**, pp. 559–65.

Wang, N. and Ingber, D.E. (1994), 'Control of cytoskeletal mechanics by extracellular matrix, cell shape and mechanical tension', *Biophysical Journal*, **66** (6), pp. 2181–9.

Woolf, N.J., Young, S.L., Johnson, G.V.W. and Fanselow, M.S. (1994), 'Pavlovian conditioning alters cortical microtubule-associated protein-2', *NeuroReport*, **5**, pp. 1045–8.

Wulf, R.J. and Featherstone, R.M. (1957), 'A correlation of van der Waals constants with anesthetic potency', *Anesthesiology*, **18**, pp. 97–105.

Yu, W. and Baas, P.W. (1994), 'Changes in microtubule number and length during axon differentiation', *J. Neuroscience*, **14** (5), pp. 2818–29.

237

NON-LOCALITY AND OBJECTIVITY IN QUANTUM STATE REDUCTION

Quantum coherence and reality: In celebration of the 60th birthday of Yakir Aharonov. Int. Conf. Fundamental Aspects of Quantum Theory. Eds J. Anandan and J. L. Safko. World Scientific, Singapore, 1994

In honour of Yakir Aharonov's 60th birthday. Once more, issues that have been treated before (e.g. Chapters 227 and 218) are treated again, but here the questions of locality and energy conservation in the (Diósi–Penrose) state-reduction scheme are explored.

522 Non-locality and objectivity in quantum state reduction. *Quantum coherence and reality: In celebration of the 60th birthday of Yakir Aharonov. Int. Conf. Fundamental Aspects of Quantum Theory.* Eds J. Anandan and J. L. Safko. 1994

Non-Locality and Objectivity in Quantum State Reduction

Roger Penrose
Mathematical Institute, Oxford, UK

Abstract

An example of quantum non-locality is presented ("magic dodecahedra") which illustrates Bell's theorem without probabilities. A scheme is then put forward for the objective reduction of the quantum state-vector when too large displacements of mass are involved in a superposition between two quantum states. In this scheme, the reduction time is roughly the reciprocal of the gravitational self-energy of the difference of the two mass distributions, measured in absolute units.

It is a pleasure to be able to pay my respects to Yakir Aharonov in honour of his 60th birthday. I shall briefly describe two ideas that have to do with that subject - *quantum mechanics* - which has engaged so much of his attention, and to which he has made so many surprising and profound contributions. The first is an example that illustrates one of the theory's most puzzling features: quantum (Bell) non-locality, without probabilities. The second represents a new angle on the measurement problem.

1 Magic Dodecahedra

I have described this non-locality example several times elsewhere [1,2,3] so it will not be necessary to give more than a very brief outline of what is involved. The system consists of two atoms of spin 3/2 which are initially produced in a combined state of spin 0 and then slowly separated to a great distance from one another without disturbing their individual spins. Measurements are subsequently made on the two atoms individually, each measurement being of a particular yes/no kind and corresponds to one of 20 possible directions in space: those which are represented by the vertices of a regular dodecahedron, as measured out from the centre. Thus, we imagine two widely separated but parallel-oriented regular dodecahedra (which, for dramatic effect, we can be imagined as being here on earth and on a planet orbiting α-Centuri, respectively), each of which has a spin 3/2 atom at its centre. Each measurement would be defined by choosing one of the vertices of one dodecahedron and ascertaining whether the amount of its central atom's spin in that direction - i.e. the m-value, in that direction - has the particular value 1/2. If this is found to be so, then this is the answer "yes"; and one envisages that a bell rings, indicating that the measurements on that particular atom have come to an end. If the value 1/2 is not obtained ("no"), then the three possible $m \neq 1/2$ states (namely 3/2, -1/2, -3/2) are combined without disturbing the phase relations between them, and the measurement is repeated in some other direction.

For example, we could envisage performing the measurement with a Stern-Gerlach type of apparatus, oriented appropriately in the chosen direction, and where only *one* of the four different beams (the one corresponding to $m=1/2$) is examined, yielding the "yes" answer (bell ringing) if the atom is found in that beam. Otherwise, by appropriate reversing of the magnetic fields, the three remaining beams are brought together without disturbing their relative phases. An exactly similar spin measurement is then performed in some other direction, corresponding to a different vertex of the dodecahadron, and so on.

There are just two different properties that we shall need, concerning the results of the joint measurements of this type that can be performed on the two atoms - by myself here on earth and by my colleague Alfie St.Uri, on α-Centuri. These concern sequences of measurements of the following type. One of the vertices of the dodecahedron is singled out - call this vertex the SELECTED one - and measurements are performed corresponding to the three vertices of the dodecahedron that are *adjacent* to the SELECTED one (but not in the direction of the SELECTED one itself). It may be ascertained that the "yes" eigenstates of these three measurements are all *orthogonal* to one another, and it follows that the three measurements necessarily *commute* - so it makes no difference in which order the three are performed. We deduce the first of the two properties that we shall need:

> (1) If Uri and I happen to SELECT *diametrically opposite* vertices on our respective dodecahedra, then the bell rings for one of my measurements if and only if it rings for Uri's diametrically opposite measurement, this being irrespective of whether it rings on the first, second, or third of the measurements adjacent to our SELECTED vertices.

The second property is a little harder to ascertain, although this can be done without further explicit calculation (cf. ref. 1 for details):

> (2) If Uri and I happen to SELECT *corresponding* vertices on our respective dodecahedra, then the bell must ring for at least one of the six measurements that we propose to make.

If we are to assume that the bell-ringings are determined according to some kind of local hidden variable theory - or, simply, that what happens on α-Centuri is completely independent, in the ordinary classical way, of the measurements that I choose to make here on earth, then we can quickly deduce three things concerning the results of measurements on my own dodecahedron alone:

> (a) Each vertex of my dodecahedron is preassigned as either a *bell ringer* (colour it WHITE) or as *silent* (colour it BLACK), irrespective of the ordering in which the measurements are made adjacent to any SELECTED vertex.
> (b) No two next-to-adjacent vertices can be both bell-ringers (WHITE).
> (c) The six vertices adjacent to one or other of a pair of opposite vertices cannot be all silent (BLACK).

It is a nice combinatorial exercise to show that no colouring of the vertices of a dodecahedron WHITE or BLACK is possible, according to the rules (b) and (c). This shows that the assumption of classical independence between the atom here on earth and the atom on α-Centuri must be false - assuming that the expectations of standard quantum theory are maintained. In Einstein's terminology, there is a "spooky action at a distance" between the results of the measurements on the spin 3/2 atoms that Uri and I might choose to make. For earlier examples of Bell non-locality without probabilities see[4–9], and the review article by Brown[10]; also [11–13] for results that can be adapted to give non-local examples of this nature.

2 The Role of Gravity in Quantum State Reduction

I have frequently argued the case that the phenomenon of state-vector reduction must be a *real* physical effect of some kind, and not just an illusion, or a property of conscious observers, or just some tricky matter of finding the right "interpretation" of the quantum formalism. Moreover, I have maintained that the physics that is involved must be something in which the effects of *gravity* are crucial (cf. also [14–17]). Of course, it is clear that there are many differing viewpoints with regard to this phenomenon, and I shall certainly make no serious attempt to convert anybody. The motivations that underlie my own approach are various, but I believe that a number of independent arguments can be given in favour of a fundamental role for gravity in quantum state reduction .

For me, one of the strongest comes from the study of the *space-time singularities* in the big bang and black holes. As a fundamental ingredient to the second law of thermodynamics, it is necessary that the big bang's singularity must have been enormously constrained - to such an extraordinary precision that only one part in (at least) $10^{10^{123}}$ of the available phase space was made use of. (Very likely, the precision is considerably greater than this, depending upon the actual baryon content of the universe the precision being infinite for a spatially infinite universe.

This figure is calculated on the basis of the Bekenstein- Hawking formula for black hole entropy, assuming a total baryon content of about 10^{80}. The necessity of this kind of precision is not removed by inflation[18].) It is this enormous constraint on the gravitational degrees of freedom in the early universe, together with the fact that the singularities of black holes - or of an all-embracing big crunch - seem to be subject to no constraint at all, that gives us the powerful second law of thermodynamics in the form that we know it. The structure of space-time singularities is generally accepted to be a quantum gravity effect - or at least an effect of whatever the correct union of quantum theory with gravitational theory might be. Since the initial and final types of singularity seem to need to have such grossly different structures, this strongly indicates that whatever this quantum-gravity union might turn out to be, it must be a time-asymmetric theory. The indications are, therefore, that something more than just the standard time-symmetric procedures of unitary evolution must be involved; the time *a*symmetric phenomenon of quantum state reduction must also be part of this entire unified picture.

It is possible to be more explicit about this link between the time-asymmetry of space-time singularities and that of the reduction procedure[19,20], but I have no wish to repeat the arguments in detail here. The essential point is that the complicated high-entropy singularities of gravitational collapse serve to "absorb information", causing an effective *reduction* in phase-space volume. Over the totality of all possible states, this must be precisely balanced by a corresponding effective *increase* in phase-space volume that results from an indeterminacy in the evolution of physical systems. This indeterminacy is argued to be that which is arises in state-vector reduction. (The phase-space volume increases because, in effect, when the state gets reduced there are several different alternative outputs for each input; whereas, given the output, there is generally only one plausible input that need be considered.) It is this necessary balance between these two seemingly disparate parts of physics that tells us that these two parts of physics must actually be one and the same. Thus, not only is the structure of space-time singularities - and consequently also the second law of thermodynamics - a quantum gravity effect, but so also must be the very process of quantum state reduction.

I have argued earlier that state reduction should be something that comes about when space-times would have to be superposed which differ "too much" from one another, in the sense that the difference between the space-times is of the order of "one graviton" or more (so nature abhors superpositions between sufficiently different space-time geometries). I have now come to the conclusion that we should not regard this measure of difference as representing something absolute - for which quantum linear superpositions would be forbidden whenever this value is exceeded. Rather, we should consider that there is a *rate* at which reduction occurs, this rate being large for space-times that differ by a large amount, and small, when the space-times do not differ much. Thus there is to be an instability involved in space-time superpositions, giving a kind of *half-life* for the superposed state, that is

CHAPTER 237

of the order of the reciprocal of the appropriate measure of the difference between the superposed states.

What is this measure of difference? In a recent article[21], I gave some rather tenuous motivations for a measure of difference between the two weak quasi-static gravitational fields that are associated with two different Newtonian mass distributions. This can be reformulated as the *gravitational self-energy* of the *difference* between the two mass distributions. The present proposal, then, is to take this self-energy E, measured in Planckian units (i.e. the *absolute* units for which $G = c = \hbar = 1$). Then E^{-1} gives something, in Planckian units, that is of the order of the *time* that the superposition persists before it reduces to that given by either one mass distribution or the other. That is to say, E^{-1} is roughly a half-life for the superposed state to decay into one state or the other.

There is a particular advantage in a viewpoint of this nature that is not shared by most other proposals for "realistic" quantum state reduction (such as that of Ghirardi, Rimini, and Weber[22]). In the case of an unstable particle, there is always an uncertainty in the mass of the particle, this uncertainty (in units for which $\hbar = c = 1$) being of the order of the reciprocal of the lifetime. Thus, for any state-reduction process of the general kind that I am proposing here, we expect some kind of mass-energy uncertainty that is inherent in the superposed state. With the present proposal, this uncertainty would have to be of the same order as the self-energy in the gravitational field of the difference between the two mass distributions under consideration. This self-energy, according to classical general relativity, is not well defined - or, at least, it is not localizable - in a coordinate-independent way.

In classical general relativity, this is an inherent feature of the theory. The quantity T_{ab} that occurs in Einstein's equation $R_{ab} - 1/2Rg_{ab} = 8\pi T_{ab}$ describes all the energy of *matter*, but it does not directly take into account the energy in the gravitational field. That energy is non-local, and cannot be meaningfully assigned a local measure of density. Thus, there is no tensor quantity, independent of coordinate choice, which describes this energy. Nevertheless, gravitational field energy is "real", in the sense that it must be taken into account in physical processes, such as the (positive) energy that is carried away in the form of gravitational radiation from a double neutron star system, or the (negative) contribution to the total mass of a celestial body, such as Jupiter, owing to its gravitational self-energy. However, the mass-energy in gravitation is a fundamentally slippery quantity, which cannot be meaningfully localized.

A feature of the present proposal for quantum state reduction is to take advantage of this slipperiness in order to evade an energy problem that seems to be an essential feature of any model of state reduction in which that process is taken to be a "real" phenomenon. Basically, if "quantum jumps" are taken to be *real*, then the mass-energy distribution in a system undergoes local violations of energy conservation when the jumps occur. In the original reduction scheme due to Ghirardi, Rimini, and Weber[22] (GRW), for example, there is a small energy violation involved in the "hits" that effect the reduction process. One of the physical ingredients of the present scheme is that there is the potential possibility of dovetailing one of these energy problems with the other - that of classical general relativity with that of quantum state reduction - so that a consistent overall scheme may be obtained. At the time of writing, however, I have not worked out the details of how this dovetailing is fully to take place. It should be pointed out, moreover, that the gravitational variant of the GRW scheme put forward by Diósi[17] has considerably more serious energy problems than the original GRW scheme - to the extent that it is in gross conflict with observation, as was pointed out by Ghirardi, Grassi, and Rimini[23]. These three authors suggested a modification that removed this observational conflict, but at the expense of introducing an *ad hoc* parameter that was not present in Diósi's proposal. The present proposal differs from that of Diósi, though it has a number of features in common with it. More details will need to be sorted out before it can be ascertained whether, within the scheme of ideas that I am setting forth here, one can construct a detailed proposal that is completely free of such energy problems.

In the meantime, we can at least examine the orders of magnitude that arise with the present scheme. Let us first take note the values of some of the standard physical units in terms of the dimensionless Planckian ones (for which $G = c = \hbar = 1$):

$$\begin{gathered}\text{second} = 1.9 \times 10^{43},\ \text{day} = 1.6 \times 10^{46},\ \text{year} = 5.9 \times 10^{50},\\ \text{metre} = 6.3 \times 10^{34},\ \text{cm} = 6.3 \times 10^{32},\ \text{micron} = 6.3 \times 10^{28},\\ \text{radius of nucleon} = 10^{19},\ \text{mass of nucleon} = 10^{-19},\\ \text{gram} = 4.6 \times 10^{4},\ \text{erg} = 5.2 \times 10^{-17},\ \text{degree Kelvin} = 4 \times 10^{-33},\\ \text{density of water} = 2 \times 10^{-94}\ .\end{gathered}$$

If we consider a uniform sphere of radius a and mass m whose state is gradually evolved into a superposed state of two different locations, separated from one another by a distance comparable with their radius a, then we find a gravitational self-energy E, for the difference between the two mass distributions, which is the order of

$$E = m^2/a.$$

Thus, according to the proposal I am putting forward, this superposed state is unstable, and would decay into the state in which the sphere is *either* in one location *or* in the other, in a time (half-life) that is of the general order of

$$T = a/m^2.$$

In terms of the density ρ (assumed uniform) and radius a of the sphere, this is very roughly

$$T = 1/(10\rho^2 a^5).$$

Now consider various examples. For a nucleon, assuming its radius to be its Compton radius

$$\begin{aligned} T &= 10^{19}/10^{-38} = 10^{57} \\ &> \text{million years.} \end{aligned}$$

Since ordinary laboratory experiments take place in time scales that are much less than this, there is - according to the present scheme - no danger of any discrepancy with the predictions of standard quantum theory for a quantum system consisting of just a few nuclear particles. In particular, the results of neutron interference experiments are not contradicted. For a droplet of water, of radius a, we find, approximately,

$$T = 10^{189} a^{-5}$$

so we get, very roughly,

$$\begin{aligned} T = 10 \text{ days}, &\text{ if } a = 10^{-5} \text{ cm} \\ T = 10^{-1} \text{ seconds}, &\text{ if } a = 10^{-4} \text{ cm} = 1 \text{ micron} \\ T = 10^{-6} \text{ seconds}, &\text{ if } a = 10^{-3} \text{ cm}. \end{aligned}$$

Thus we see that, in a sense, a "turnover" from quantum to classical behaviour occurs at roughly a micron's scale. These figures, and other related ones, are not at all unreasonable. They do not seem to contradict anything obvious about quantum or classical behaviour.

In making this statement, I am taking into account two factors that I have not mentioned so far. The first is that in the above estimates I have treated the bodies as *uniform* objects, and not as composed of individual atomic particles. In fact, in any ordinary situation of a superposition between two mass distributions, one of which is a rigid translation of the other, the granular (atomic) nature of the distributions turns out not to be important for the calculation of the reduction rate. But we can also consider a *different* type of situation: where we have a superposition of two mass distributions which do not differ macroscopically at all. Instead, it is now to be their submicroscopic constituents that are located differently in the two states. Depending upon the nature and the amount of movement of these constituents, we find in this situation that, for an equal reduction rate, the total size of the material that is involved in the superposition would tend to be somewhat larger - but not enormously larger - than that considered above.

Secondly, these considerations are important when reduction occurs because the *environment* becomes entangled with the quantum system that is under study. Indeed, according to the present proposal, in any practical situation that one can easily envisage at the moment, it would indeed be the disturbed environment that effects the reduction. Thus we obtain nothing different from the conventional picture of state-vector reduction, in which it is the "decoherence" caused by the environment that causes the reduction - except that now the reduction must be considered as a real physical effect, not just something that takes place "for all practical purposes" (John Bell's "FAPP"). It would, no doubt, take a delicately organized experimental set-up to detect any differences between the present proposal and conventional quantum mechanics. Nevertheless, differences would be detectable in principle. One would need to arrange things so that some "large" quantum system can remain isolated from its surroundings for sufficiently long that, according to the present scheme, state-reduction should take place spontaneously within that time-scale, entailing a loss of phase coherence. On the other hand, standard quantum mechanics would demand that such coherence would be maintained for as long as the system remains isolated.

Finally, it should be remarked that these considerations leave us a long way from an actual *theory* of gravitationally-induced state-vector reduction. The difficulties of providing a coherent picture of the reduction in accordance with the principles of relativity are well known, and were stressed many years ago by Yakir and his colleagues[24].

Acknowledgements

I am grateful to Abhay Ashtekar, Ted Newman, and Lenny Susskind for illuminating suggestions; I also thank NSF contract PHY-9396246 at Syracuse and Penn State Universities for partial funding assistence.

References

1. R. Penrose *On Bell non-locality without probabilities: some curious geometry*, in *Quantum Relections* (in honour of J.S. Bell), eds. J Ellis and D. Amati (Cambridge Univ. Press, Cambridge 1994 to appear)
2. R. Penrose in *The renaissance of General Relativity* (in honour of D.W. Sciama), eds. G. Ellis, A. Lanza and I. Miller (Cambridge Univ. Press, Cambridge 1993)
3. J. Zimba and R. Penrose *Stud. Hist. Phil. Sci.* (1993)
4. P. Haywood and M.L.G. Redhead *Found. Phys.* 13 (1983) 481-99

5. A. Stairs, in *Philos. of Sci.* 50 (1983) 578-602
6. H.R. Brown and G. Svetlichny *Found. Phys.* 20, (1990) 1379- 87.
7. D.M. Greenberger, M.A. Horne and A. Zeilinger *Bell's Theorem, Quantum Theory, and Conceptions of the Universe*, ed M. Kafatos (Kluwer Academic, Dordrecht, The Netherlands 1989), pp. 73-76.
8. D.M. Greenberger, M.A. Horne, A. Shimony and A. Zeilinger *Am. J. Phys.* 58, (1990) 1131-43.
9. R.K. Clifton, M.L.G. Redhead and J.N. Butterfield *Found. Phys.* 21, (1991) 149-84; errata, *Found. Phys. Lett.* 4, (1991).
10. H.R. Brown in *Bell's Theorem and Theorem and the Foundations of Physics* eds. A. Van der Merwe and F. Selleri (World Scientific, Singapore 1993)
11. J.S. Bell *Revs. Mod. Phys.* 38, (1966) 447-52.
12. S. Kochen and E.P. Specker *J. Math. Mech.* 17 (1967) 59-88.
13. A. Peres *J. Phys.* A24 (1991) L175-8.
14. F. Károlyházy *Nuovo Cim.* A42, (1966), 390
15. A.B. Komar *Int. J. Theor. Phys.* 2, (1969) 157-60.
16. F. Károlyházy, A. Frenkel and B. Lukács in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford 1986).
17. L. Diósi *Phys. Rev.* A40, 1165 (1989).
18. R. Penrose in *Fourteenth Texas Symposium on Relativistic Astrophysics* ed. E.J. Fenyves *N.Y. Acad. Sci., New York* 57 (1989) 249- 64.
19. R. Penrose *Quantum Concepts in Space and Time* eds. R.Penrose and C. J. Isham, Oxford University Press 1986), pp. 129-146
20. R. Penrose *The Emperor's New Mind* (Oxford University Press (1989))
21. R. Penrose, in *General Relativity and Gravitation 1992: Part 1, Plenary Lectures* eds. R.J. Gleiser, C. Kozameh and O.M. Moreschi (I.O.P. Publ. 1993) pp. 179-89.
22. G.C. Ghirardi, A. Rimini and T. Weber *Phys. Rev.*, D34, (1986) 470.
23. G.C. Ghirardi, R. Grassi and A. Rimini *Phys. Rev.* A42 (1990) 1057.
24. Y. Aharonov and D.Z. Albert *Phys. Rev.* D24 (1981) 359.

238

MUST MATHEMATICAL PHYSICS BE REDUCTIONIST?

Nature's imagination. Ed. J. Cornwell, 12–26. Oxford University Press, 1995

CHAPTER TWO

Must mathematical physics be reductionist?

ROGER PENROSE

I think that one of the things to be learned from Freeman Dyson's presentation is that one cannot be a self-respecting scientist without also being a rebel. This would seem to imply that in this book we are urged to rebel against the point of view of reductionism. This is difficult to do unless we know first what reductionism is; and in fact I am not clear that I do know what the word means. Some forms of reductionism are almost synonymous with 'scientific', in which case we find ourselves driven to a contradiction. If we are going to be scientific, we have to be rebellious, but that rebelliousness leads us to resist the critical stance of the symposium—which would mean science itself if we take 'reductionist' to mean merely 'scientific'.

I am going to interpret the word 'reductionism' in various ways that are different from this. I have certainly been accused of being non-reductionist, and I have to try to understand what this charge means. There are perhaps two main alternative interpretations that come to mind, and I shall try to illustrate them in this contribution.

There are indeed certain areas of mathematics that are, in one real sense, non-reductionist. This does not prevent us from discussing problems in those areas in a very clear and characteristically mathematical way. We might take reductionism to mean the breaking-down of things into smaller and smaller parts, so that if you understand how the small parts work, that will in principle tell you how the big thing works. It may be that you need other kinds of ideas to say interesting things about the big things, as Freeman Dyson stressed, but here, the idea of reductionism is that the behaviour of the big things we study is governed by the behaviour of the individual units of which they are composed.

That is one strand of reductionism. Another strand of reductionism may be related to this by asking if knowledge of the present behaviour of those small units will allow us to predict the behaviour of the big things in the future? Determinism has been a central concern of philosophy and physics for many centuries: we may understand it as the proposition that the behaviour of the system at one time will determine what it does at a later time. It is important to emphasize that the *computability* of the behaviour at later times is a separate question. I want to stress the difference between computability and determinism because the distinction may prove to be an important one. Before we address it, I should like to return to our first interpretation of reductionism and illustrate how it is possible to talk about holistic concepts in a clear and mathematically precise fashion.

HOLISM AS MATHEMATICS

Let us start by investigating what we mean by a holistic concept. Figure 2.1 illustrates an impossible triangle. By virtue of what particular property is the triangle impossible? The fact that it possesses this feature of impossibility is clear. It is supposed to be an image that conveys to the mind a particular three-dimensional object, but that three-dimensional object simply cannot exist in ordinary space. We must ask ourselves what is wrong with the picture: can we point to somewhere in the picture where the mistake was made?

We might locate the impossibility in one specified corner of the triangle, so that if we covered up that corner the figure would make sense as the

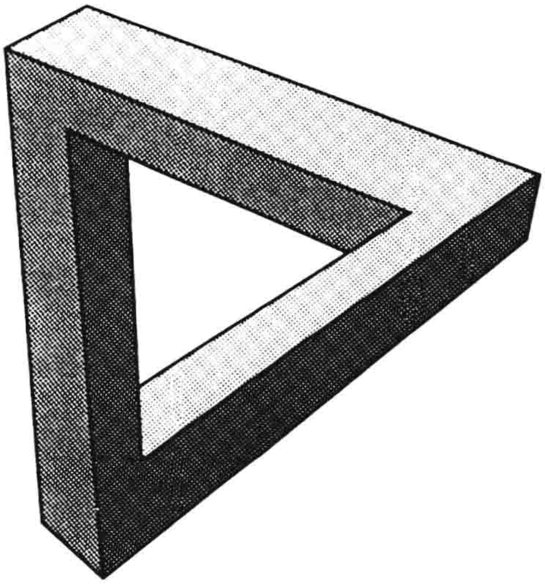


Fig. 2.1. An impossible triangle. Where is the impossibility?

representation of a possible three-dimensional object. We could then say that the impossibility has disappeared. Apparently it did not reside in the rest of the picture. If we hide *any* one corner, or remove *any* edge, the structure suddenly becomes possible. When we study the complete picture, however, the impossibility re-emerges.

This impossibility is a property of the whole structure. It cannot be localized in some part: it is a feature of the complete mathematical object, so it is a *holistic* property. There are areas of mathematics dedicated to discussing this kind of property in a rigorous way.

Let us imagine breaking the figure into three pieces, as illustrated (Fig. 2.2), and then glueing the parts together again. We could continue breaking the structure down, and each of the pieces would be a possible object in ordinary space. The process of glueing them together is the operation which eventually produces the impossible triangle. In technical mathematical terms, we take all the pieces together with all the specific glueing operations and by a factoring-out procedure, we extract what is called the cohomology. This mathematical concept abstracts what it is that interests us: the measure of impossibility of the triangle. In this particular instance, we can give this measure as a single real number, describing the 'degree of impossibility' of the triangle.

If I break this structure somewhere, so that it becomes a possible three-dimensional object, I can measure the ratio of the distances from my eye of one end of the break to the other end of the break. That ratio is a measure of the

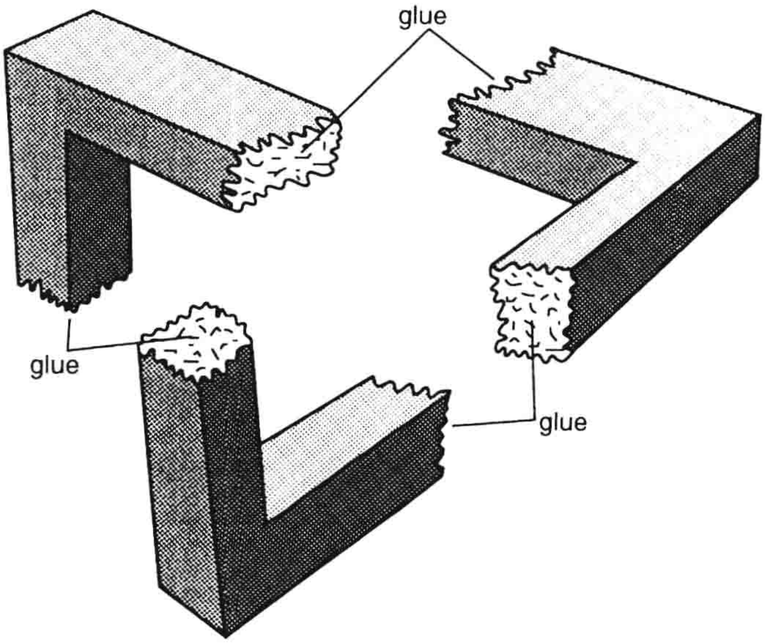


Fig. 2.2. Cohomology: an impossible triangle broken and reassembled.

impossibility of the structure and it will be the same, no matter where I cut the figure. This holds also with more than one break, no matter how many pieces I reduce it into, provided that I multiply all the relevant factors together. That perhaps gives you something of the idea of cohomology. In this particular structure, cohomology supplies the degree of impossibility of the object, and it is a perfectly well-defined and precise holistic mathematical concept.

If we take a knot (Fig. 2.3), we might ask another question: where does the property of knottedness reside in this knot? Like the impossibility of the triangle, we find that the 'knottedness' cannot be localized. It is a property of the structure as a whole. There are profound mathematical theories that deal with this question of knottedness: and though it looks simple enough, knottedness is more difficult to deal with than cohomology. Research in this field dates from the turn of the century and more recently and is now very well developed.

Another example is a Möbius band (Fig. 2.4). Where, we might ask, is the twist? If we take a piece of paper, twist it once and glue it together, we get a Möbius band. Then wherever we cut it, the twist is gone. It does not matter

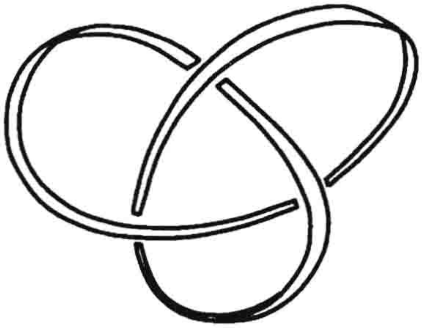


Fig. 2.3. Where does the knottedness reside in the knot?

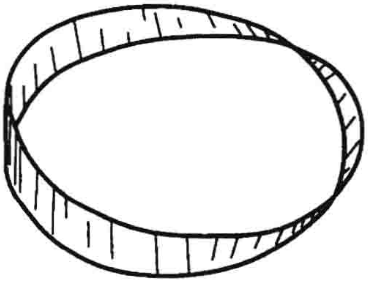


Fig. 2.4. A Möbius band. Where is the twist?

where we cut it, because once again the twist is a feature of the global structure and not something that can be localized. Once again, mathematics offers quite involved theories arising from the study of these twisted structures (called fibre bundles).

Finally we have (Fig. 2.5) a surface with a certain number of handles attached to it. Two such surfaces are said to be topologically equivalent if they can be deformed continuously one into the other. Where, we might ask, is the topology? We find that we cannot point to anywhere on the surface where we can localize it. The topology is a property of the global structure.

All these figures represent concepts important to modern physics. The idea that particles might be knots in some way goes back to Lord Kelvin. Such ideas have been revived and modernized and remain prominent in current thinking. Fibre bundles are crucial to the modern gauge theories of particle interactions. Surfaces like the one we examined are central to string theory.

Can we describe this sort of thing mathematically? There are in fact well-defined and often very subtle mathematical procedures for discussing holistic properties of the kind that I have illustrated above. There is nothing mystical about them. Holism is an entirely respectable, perfectly clear concept within modern mathematics.

QUANTUM NONLOCALITY

The objects we looked at above were mathematical, but I pointed out that they played an important role in some theories of particle physics. Moreover, holism

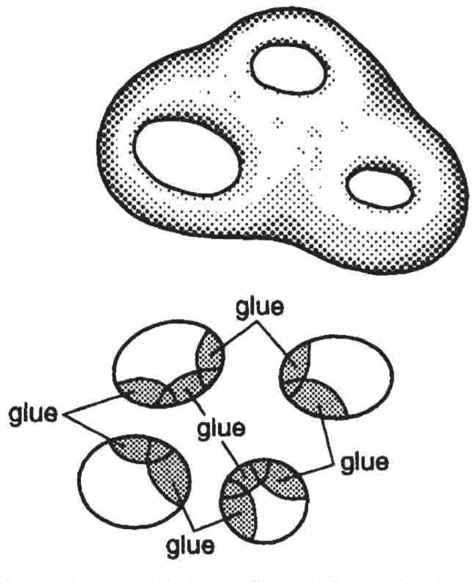


Fig. 2.5. A surface with handles. Where is the topology?

has a recognized place even in classical physics, which accepts that certain concepts are not localizable. The concept of energy in Einstein's general relativity theory, in particular, turns out to be non-local. It is the localization of the *gravitational* contribution to the energy in a system that turns out to be impossible. In general relativity, the energy and therefore (by Einstein's $E=mc^2$) also the *mass* of a system is not something we can pin down at any particular location. It has already become a holistic concept. With quantum mechanics, physics becomes holistic in an even more mysterious way.

Let us consider the phenomenon known as quantum entanglement. It was Schrödinger who first identified it, but Einstein, Podolsky, and Rosen pointed out the absurd nature of some of its implications—called EPR phenomena—and it was John Bell who showed, in effect, that quantum entanglement could be tested experimentally. The most famous such experiment was performed by Alain Aspect and his colleagues in Paris. We can describe it roughly in the following way. We have a number of atoms in one energy level, which from time to time drop down to a lower energy level. In the process, each emits a pair of photons. These photons go off in opposite directions and can be received by detectors, some 12 metres apart, that measure the photons' polarizations. The decision as to which of two possible directions is chosen in which to measure each photon's polarization is not made until they are in flight. Bell showed that any scheme in which these two photons are treated as separate independent objects would necessarily involve probabilities that satisfied certain laws. Quantum mechanics, by contrast, tells us that the probabilities will *not* satisfy such laws. The results of the experiment violated those laws in the way that quantum mechanics had predicted: so the quantum laws were vindicated. A quantum system like this, with two photons travelling a distance of 12 metres, reveals that the pair of photons cannot be regarded as consisting of two independent, separate things. This underwrites the essentially holistic nature of quantum systems. We usually cannot break down such systems into smaller and smaller units. The assumption that the photons in this experiment are independent units is not borne out by the results.

These experiments depend upon the measurement of joint probabilities. However, there has been interesting recent work of a similar nature, but which does *not* involve any probabilities (other than 0 or 1). This new work is associated with the names of Greenberger, Zeilinger, Horne, Shimony, Redhead, Clifton, and Butterfield. It concerns experiments involving three separated units, or perhaps four, instead of two. Now there are no probabilities at all: the puzzling EPR effects are obtained in simple yes/no constraints. If the objects were separate and independent entities, one would have to say that certain things could not happen, but quantum mechanics says that they *do* happen. The clear expectations are that it is quantum mechanics which is true, at least at this level, and the 'common-sense' view that widely separated objects must behave independently which is false.

There is a particular instance of Bell nonlocality without probability that I should like to discuss: I call it the *magic dodecahedra*. What we have to imagine is that here on Earth I receive a parcel sent from the far distant place where these things are produced, and an identical parcel is sent to a colleague of mine on Alpha Centauri. The package containing each magic dodecahedron comes with instructions telling us how these things behave. I should point out at this stage that the manufacturers of the dodecahedra—on the distant star Betelgeuse—have been producing these things for millions of years and they have never made a mistake. We can rely absolutely on the claims of the manufacturers when we perform our experiment.

Each of us selects a vertex of the dodecahedra and then takes three vertices adjacent to the one selected, as illustrated in Fig. 2.6. There are buttons on each of the vertices, and we press these three buttons in random order. A bell may ring; but it is going to ring on only one of these vertices and it may not ring on any of them. In the instructions I received with the dodecahedron I was told that certain things never occur, and these instructions relate what happens with my dodecahedron to what happens with the one received by my colleague on Alpha Centauri. First, if the bell rings on one of the vertices of my dodecahedron, then the bell cannot fail to ring on the opposite vertex on the other dodecahedron if my colleague presses that one, and vice versa. Thus, if my colleague happens to press a vertex on his dodecahedron directly opposite the button I press, whether it is the first, the second, or the third of the three adjacent to my selected one, then if the bell has rung for me, it must also ring for him and vice versa. Secondly, if he happens to select the same vertex as I do, and presses the three buttons adjacent to it, the instructions tell me that it

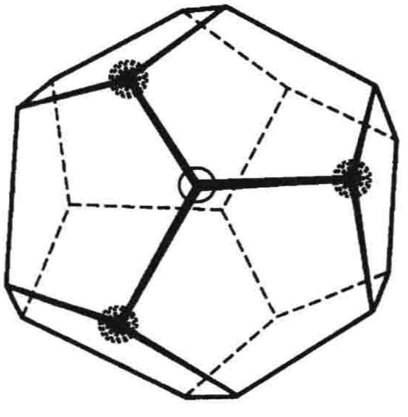


Fig. 2.6. A magic dodecahedron. The ringed vertex is selected and the three dotted vertices are pressed.

cannot be the case that the bell fails to ring for those three buttons here and also for the corresponding three on Alpha Centauri. Of course, I shall not know what has happened with his dodecahedron for a while, because my colleague is four light years away from Earth. However, since the makers of these dodecahedra have not been wrong about their behaviour in a million years, we must presumably just accept their rules.

The only problem is that there turns out to be something rather contradictory in what they say. The first rule states that if I press a button on the corner of my dodecahedron here on Earth, and the bell rings, it must then ring for the opposite button on the dodecahedron on Alpha Centauri. The makers on Betelgeuse are not to know whether my colleague on Alpha Centauri is going to choose that opposite vertex as one of the combination of buttons he presses. If my bell rings, his rings, and if my bell does not ring, his does not. The ringing or otherwise of my bell seems to be something that my dodecahedron must be prepared to do ahead of time, since the manufacturers cannot guarantee that my colleague on Alpha Centauri will not press the button opposite to mine. Our instructions require that both dodecahedra are to be carefully orientated so that the vertices can be related to each other, and it seems that my dodecahedron knows in advance whether or not to ring the bell if I press a particular button. I deduce that if I press two next-to-adjacent buttons, the bell cannot ring on both of them. Secondly, if it does not ring for the three buttons adjacent to my selected one—i.e. on three mutually next-to-adjacent ones—then it must ring on one of the opposite three buttons (see Fig. 2.5).

Taking these two properties which we seem to have conclusively derived, and doing a little work, we actually find that it is *impossible* to designate all the vertices of this dodecahedron as bell-ringers or silent, in accordance with the original guarantees of the makers. It just cannot be done. We might react by suggesting that the makers on Betelgeuse have finally made a mistake after a million years, but if we wait four light years for the message from my colleague on Alpha Centauri, we shall find that his results *do* agree with the guarantees we originally received. The magic dodecahedra are actually a curious example of *quantum entanglement*. In the above reasoning, I made the seemingly unexceptionable assumption that the two dodecahedra were *independent* objects. This assumption was false. Quantum entanglement—a genuinely holistic feature of quantum systems—entails that the objects in question are actually mysteriously connected, no matter how far apart they might be.

I shall not explain in detail how these dodecahedra are built: that is the province of those who know the appropriate physics. What I can say is that on Betelgeuse they had a spin-0 system which was split into two atoms, each of spin-$\frac{3}{2}$. At the centre of each dodecahedron is placed one of these atoms of

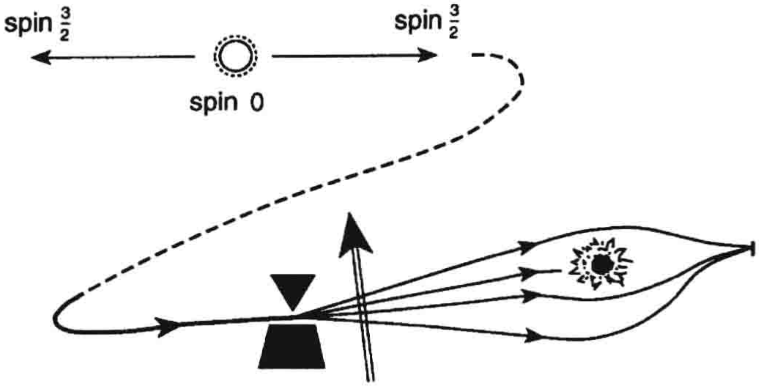


Fig. 2.7. A spin-0 system that is split into two atoms, each of spin-$\frac{3}{2}$.

spin-$\frac{3}{2}$, and if I press a button on one of the vertices, that effects a measurement in the direction of that vertex. There are four possible outcomes. The atom chooses one of those four, and if it is in the second from top (to the experts: $m = \frac{1}{2}$) the bell rings (see Fig. 2.7). If not, the bell does not ring, in which case the remaining three possibilities are reassembled and we try again in another direction. With present technology, it might even be possible to set this up and perform an experiment of this nature.

According to standard theory and observed experimental fact, quantum entanglement certainly does occur. Widely separated physical objects are not independent of one another. I should also emphasize that according to standard quantum-mechanical theory, one cannot use this absence of independence to send a message faster than the speed of light from one object to the other. Nevertheless, the way in which the dodecahedra behave cannot be understood if we insist on regarding them as separate and independent objects. This is indeed an instance of the phenomenon of quantum entanglement. The phenomenon generally is prominent in many quantum systems. This particular example neatly illustrates that physical behaviour cannot be explained if we insist on treating the units involved as separate, independent local objects.

Why isn't everything entangled? If these two particles, four light years away from each other are entangled, why isn't the whole universe an extraordinary entangled mess? How can we deduce anything which has the semblance of locality? After all, if this 'local' form of reductionism is not right, and the universe cannot be separated into these independent pieces, why does our physics work so well?

The answer is not altogether clear, but it is noteworthy that in our example of the magic dodecahedra there was a process of *measurement*. Pressing the button carried out a measurement, and measurement seems to disentangle.

Once you perform a measurement you seem to have released this unit from its partner in some mysterious way. Perhaps we should ask what a measurement is, in physical terms.

Quantum theory is all about measurements but it never clarifies what a measurement actually is. In some sense, a measurement involves magnifying a quantum effect to something on a microscopic scale, so that we can see what is going on in the system. At the quantum level you might have to superpose all the possible things that might happen; once you have magnified to a macro level, only one thing or another actually happens, and not all these strange superpositions of alternatives happen all at once, which is what we must consider occurs at the quantum level.

Some would say that measurement occurs when the quantum system becomes involved with the environment in some sufficiently random way. Although the systems do not disentangle in the strict sense, they are *in effect* disentangled—or so it is claimed. We are provisionally able to treat the system as though it were not entangled, even though, in a strict sense according to standard theory, they would always remain entangled.

My own view is that measurements really *do* disentangle quantum systems: strictly speaking, this would require a change in fundamental physical theory. It is deeply unsatisfactory to have different rules applying in physics at the quantum level and at the classical level. Atoms and molecules seem to obey one set of rules which do not obtain at the classical level. Moving from one level to the other is difficult to do intelligibly: but here we are touching upon a subject of endless controversy. I had better not press the matter further here, but just make note of my own belief in the need for a new underlying theory, and move on.

HOLISM IN BRAIN ACTION?

Are quantum effects such as nonlocality and entanglement relevant to explanations of brain action? When I wrote *The Emperor's new mind*, I found it difficult to believe that nerve transmission could somehow preserve quantum entanglements and quantum superpositions, but I could not think of anything else. Nerve activity seems to be a large-scale, classical affair where each signal disturbs the environment so much that it is hard to see how quantum entanglements and superpositions could be maintained. Since then I have learned of another much more promising possibility from Stuart Hameroff, and I should like to outline it.

There is a good deal of computational activity below the level of neurones, Hameroff argues, and he points to single-cell organisms such as amoebae and paramecia, which are capable of behaving in elaborate ways despite having no nervous system at all. They do, however, possess a structure known as a cytoskeleton, which not only governs their shape but also appears to control how they behave. A paramecium can swim towards its food, or away from danger. It can negotiate obstacles and even learn by experience. Clearly it is capable of complex and sophisticated activity; yet it does not possess a single neurone. If it is indeed the cytoskeleton that controls the behaviour of a paramecium, then we may expect cytoskeletons to be important in the behaviour of other cells. In fact, all eukaryotic cells possess cytoskeletons—including individual neurones.

If we study the structure of a cytoskeleton, we find tube-like threads known as microtubules, made out of arrays of protein molecules called tubulin dimers. These structures, it seems to me, are very plausibly regarded as quantum-mechanical rather than classical-level objects. There could be effects travelling down these tubes that do not disturb the environment, and these should then be understood in quantum rather than classical terms, where quantum entanglements and superpositions are preserved. It would be difficult to tell what is actually going on within such a system, but it would be quite different from the action of a nerve. Hameroff and his colleagues believe that messages may be travelling up and down these microtubules which represent basic computational activity, and that these messages govern the behaviour of one-celled organisms. Moreover, the activity in the microtubules within neurones can apparently influence or control the synaptic connections between neurones. Though microtubules are not concerned with the actual nerve signals, they may well govern the strength of the synapses.

It seems to me to be an interesting hypothesis that collective quantum effects may be taking place in these structures. At least part of the role of the microtubules in neurones is that they are significant in *maintaining* the strength of the synapse. Without the maintenance, the strength of the connection would decay. If quantum non-local effects are important in the explanation of brain activity, this seems to be the place to start looking for them. It is worth recalling that some people have objected to any kind of quantum explanation of brain activity on the grounds that the brain is too hot for quantum effects to be important. This may be compared with the collective quantum phenomenon of superconductivity, which was originally thought to occur only at temperatures of a few degrees above absolute zero. Superconductivity is now observed at comparatively high temperatures, and I would argue by analogy that the high temperature of the brain does not rule out the occurrence of collective quantum effects that might be relevant to brain activity.

Classical physics seems incapable of explaining a phenomenon so deeply mysterious as consciousness. The physical world we know is extremely mysterious, and the more we probe its foundations the more mysterious it becomes. That mystery does not mean that the world is irrational—that it is not susceptible to scientific understanding—but it does compel us to accept mystery, and these non-local quantum phenomena are certainly mysterious.

Brain activity is often explained by drawing parallels with the operation of computers, so let us consider the possible operation of quantum computers. Hameroff and his colleagues argue that computational activity is taking place in microtubules far below the level of neurones. If this activity indeed involves quantum effects, we should investigate what the quantum-level analogue of a computer actually is. There is some theoretical work on this question: in particular, it has been shown that there are certain mathematical problems that are solved more quickly by a quantum computer (as defined by David Deutsch) than by a conventional computer—although the problems for which this is so far known to be true constitute a very special class. Quantum computers are potentially more efficient than classical computers for certain types of problem. It may be that this efficiency is an important feature of a quantum-computational behaviour in microtubules. My own purposes would require something more: a situation in which the quantum system performs actions that cannot be explained in computational terms at all.

There are indeed mathematical problems of this nature: problems that cannot be solved using ordinary computational procedures. Yet the insights that human beings have seem to lie beyond computation too, so that if we believe that what we do with our brains is rooted in physical activity, which I *do* believe, the implication is that this physical activity contains some non-computable elements. This is viewpoint C in the following table of alternatives.

Common computers, impressive and wonderful though they may be, are not sufficient to do what brains are capable of. We need something more. I believe that we need a new theoretical grasp of the relationship between quantum-level and classical physics. When we move from one level to the other, existing theory looks rather muddled. We have these things called 'measurements', which give rise to quantum *probabilities* as I discussed earlier. Somehow we change the quantum-mechanical rules when we shift from the small to the large scale when a 'measurement' is performed. The new theory that is needed to make sense of quantum theory will also be required to explain how non-computational activity can apparently arise in some physical systems—in particular, in physical brains.

Why do I say that there ought to be non-computational activity in brain action? The argument comes down to the work of Gödel, the significance of which was referred to by Freeman Dyson in his contribution. I think Gödel's

A: All thinking is computation and the mere carrying out of appropriate computations will evoke feelings of conscious awareness.

B: It is the brain's physical action that evokes awareness; and any physical action can in principle be simulated computationally. But computational simulation by itself cannot evoke awareness.

C: Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.

D: Awareness cannot be explained in physical, computational, or any other scientific terms.

theorem is indeed one of the greatest achievements of the twentieth century. One implication of the Gödel argument is that mathematical belief cannot be encapsulated by any knowable algorithm. The 'knowable' part is the important thing here. What the argument of Gödel tells us is that we cannot know any algorithm A that we are to believe underlies our mathematical beliefs. If we know and believe A, then we must believe additionally that A is consistent. It must not give a nonsense result such as $1 = 2$, for example. If we believe that A is consistent, we must believe its Gödel statement, which we shall call $G(A)$, where $G(A)$ codes the consistency of A. $G(A)$, however, lies beyond the scope of A. The system A cannot encapsulate $G(A)$, so long as A actually is consistent. It follows that there cannot be any A that can be known to us to capture the full extent of our mathematical beliefs.

It seems to me appropriate to conclude that we do not use any definite algorithm at all—but the Gödel argument does not quite show this. What the Gödel argument does not discuss is whether we are acting in accordance with an unconscious unknowable algorithm—or at least an algorithm whose *soundness* is not knowable. Mathematicians, you might well think, should know what they are doing. But if they are in reality the servants of some hidden algorithm operating in their heads then they cannot know what this algorithm is. That they are acting in this unknowable way is a possible but, I think, an unlikely proposal. It opens up a much longer discussion. Up to this point I have employed fairly rigorous mathematics. To go beyond this, it seems that we must turn to other kinds of argument. Are we, as mathematicians, Usual form of Gödel argument

$$\text{Formal system } F \begin{cases} \text{axioms} \\ \text{rules of procedure} \end{cases}$$

can form Gödel proposition G(F).

If F is **consistent** (and broad enough for arithmetic), then

$$\mathrm{G}(F) \text{ is } \begin{cases} \text{not derivable within } F \\ \text{TRUE} \end{cases}$$

N.B. F consistent $\Leftrightarrow$ $0=1$ cannot be proved in F.

Belief

If we believe F sound
then we believe F consistent
therefore we believe G(F) TRUE

Therefore belief is non-computational.

really acting in accordance with an unconscious unknowable algorithm? One inference from such a proposal would be that the reasons we offer for believing our mathematical results are not the true reasons for such belief. Mathematics would depend upon some unknown calculational activity of which we were never aware.

Although this is possible, it seems to me unlikely as the real explanation for mathematical conviction. We have to ask ourselves how this unconscious unknowable algorithm, of value only for doing sophisticated mathematics, could have arisen by a process of natural selection. An alternative argument, which is a kind of composite of the strong artificial intelligence view and a religious one, has been brought to my attention. Its proponents would accept that while the brain may be a kind of computer, it has been programmed by 'the best programmer in the business'. *We* cannot know this mysterious algorithm, but *God* does and it was *he* who instilled it within us, rather than its having arisen by a process of natural selection.

My personal belief is that our mathematical understanding is a product of natural selection. But I would argue that this quality of 'understanding' cannot be an algorithmic thing at all. If it were algorithmic it would have to provide a formal mathematical system that was specific to doing sophisticated mathematics. Such things are very far from ordinary human experience. If the ability to perform advanced mathematics had had a strong selective advantage in the past, there might have been a straightforward case for the contrary belief; but there is no indication of any such selective advantage. There is not even an advantage now! This makes it very hard to see how an algorithm specific to sophisticated mathematics could ever be selected for in evolution. We need the selective processes to have acted in a context in which the very quality of 'understanding' itself—necessarily a non-algorithmic thing—can have had a fundamental selective advantage.

I claim that the Gödel argument demonstrates that whatever understanding is, it is indeed *not* a computational thing. This allows that natural selection could operate for this general non-computational quality—a quality which could be applied to a whole range of problems and not simply to mathematics.

If mentality is a function of brain action, and we accept that brain action is subject to the same laws of physics as everything else, those physical laws must allow room for a non-computational action. The most promising place we should explore in order to locate this action is the region between the quantum and classical levels of physics. This is an area that is ill understood at present, but I believe it is crucial to the issue of non-computational physical activity.

REFERENCES

Hameroff, S. R. (1987). *Ultimate computing*. North Holland, Amsterdam.

Penrose, R. (1989). *The Emperor's new mind*. Oxford University Press.

Penrose, R. (1991). On the cohomology of impossible figures. *Topologie Structurale*, **17**, 11–16.

Penrose, R. (1994). *Shadows of the mind*. Oxford University Press.

Redhead, M. L. G. (1987). *Incompleteness, nonlocality and realism*. Clarendon Press, Oxford.

239

CONCERNING SPACE-TIME POINTS FOR SPIN 3/2 TWISTOR SPACES

Twistor Newsletter **39** (1995)

Again this is the spin-3/2 approach to curves-space twistors (Chapter 233, etc.), but there is a new angle here, relating these things to exact sets (Chapter 114).

Concerning Space-Time Points for Spin $3/2$ Twistor Spaces

In $\mathbb{T}$N 38 (pp. 1-9) (and see that article for earlier related references) I discussed various sheaves, exact sequences, etc. relevant to the spin $3/2$ approach to the long-standing programme of defining the appropriate twistor space $\mathcal{T}$ for a general vacuum (Ricci-flat) space-time $\mathcal{M}$. Certain further ideas directed towards finding the twistors for $\mathcal{M}$ (i.e. points of $\mathcal{T}$) have arisen, but these remain inconclusive and will not be described here. Instead, I shall concern myself with a different question. Let us imagine that we have found the space $\mathcal{T}$. How would we go about re-constructing $\mathcal{M}$ from $\mathcal{T}$? In other words, what do we expect would be the definition of a space-time point (i.e. point of $\mathcal{M}$) in terms of $\mathcal{T}$?

Let us recall a way in which (complex) space-time points have been viewed in relation to the "googly" programme and non-linear graviton construction (R.P. in $\mathbb{T}$N 8, pp. 32-34). We first consider the standard (Poincaré-invariant) exact sequence for flat twistor space $\mathbb{T}$:

$$0 \to \mathbb{S}^A \xrightarrow{i} \mathbb{T}^\alpha \xrightarrow{p} \mathbb{S}_{A'} \to 0$$

where the second map (i) takes ω^A to $(\omega^A, 0)$, and the third (p) takes $(\omega^A, \pi_{A'})$ to $\pi_{A'}$. A point q in $\mathbb{C}\mathbb{M}$ can be associated with either of the two maps $q, \hat{q}$ indicated by the dotted lines in

$$\begin{tikzcd} 0 \arrow[r] & \mathbb{S}^A \arrow[r, "i"] & \mathbb{T}^\alpha \arrow[l, dotted, bend left, "\hat{q}"] \arrow[r, "p"] & \mathbb{S}_{A'} \arrow[l, dotted, bend left, "q"] \arrow[r] & 0 \end{tikzcd}$$

where $\hat{q}$ takes

$$(\omega^A, \pi_{A'}) \text{ to } \omega^A - i q^{AA'} \pi_{A'}$$

(q^a being the position vector of q), and the map q takes

$$\pi_{A'} \quad \text{to} \quad (iq^{AA'}\pi_{A'}, \pi_{A'}).$$

The composition $p \circ q$ is the identity on $\mathbb{S}_{A'}$, and $\hat{q} \circ i$ is the identity on $\mathbb{S}^A$.

In the non-linear graviton construction for an anti-self-dual complex vacuum $\mathcal{M}$, the twistor space $\mathcal{T}$ still has a canonical projection $\mathcal{T} \xrightarrow{p} \mathbb{S}_{A'}$, and space-time points (cross-sections of $\mathcal{T}$ in its capacity as a fibration over $\mathbb{S}_{A'}$) can be identified with (holomorphic) maps $q: \mathbb{S}_{A'} \rightarrow \mathcal{T}$ for which $p \circ q$ is the identity on $\mathbb{S}_{A'}$

$$\begin{tikzcd} \mathcal{T} \arrow[r, shift left, "p"] & \mathbb{S}_{A'} \arrow[l, shift left, dotted, "q"] \end{tikzcd} .$$

(More correctly, $\mathbb{S}_{A'} - \{0\}$ should replace $\mathbb{S}_{A'}$ in the preceding assertion.) In the proposals for a "googly graviton" construction, the points of a self-dual complex vacuum $\mathcal{M}$ would be intended to arise correspondingly from the twistor space $\mathcal{T}$ in accordance with

$$\begin{tikzcd} \mathbb{S}^A \arrow[r, shift left, "i"] & \mathcal{T} \arrow[l, shift left, dotted, "\hat{q}"] \end{tikzcd} .$$

It is of interest to explore the way in which space-time points might come about in relation to a "spin 3/2 definition" of the twistor space $\mathcal{T}$ for a general complex vacuum $\mathcal{M}$. Let us recall the exact commutative diagram Ⓢ appearing on p.7 of TN38, where I am also introducing additional maps $q, \hat{q}$ (to be explained shortly) in the columns (upwards)

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 & \\
0 \arrow[r] & \{\mathring{\Omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\rho}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\rho}\}/\{\mathring{\omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\alpha}\} \arrow[r] \arrow[d, shift right, "i"'] & 0 \\
0 \arrow[r] & \begin{Bmatrix}\Omega \\ \Pi\end{Bmatrix} \arrow[r] \arrow[u, shift right, dotted, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\omega \\ \pi\end{Bmatrix} \arrow[r] \arrow[u, shift right, dotted, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\rho \\ \sigma\end{Bmatrix} \arrow[r] \arrow[u, shift right, dotted, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\rho \\ \sigma\end{Bmatrix}/\begin{Bmatrix}\omega \\ \pi\end{Bmatrix} \arrow[r] \arrow[u, shift right, dotted, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\alpha \\ \beta\end{Bmatrix} \arrow[r] \arrow[u, shift right, dotted, "\hat{q}"'] \arrow[d, shift right, "p"'] & 0 \\
0 \arrow[r] & \{\Pi\} \arrow[r] \arrow[u, shift right, dotted, "q"'] \arrow[d] & \{\pi\} \arrow[r] \arrow[u, shift right, dotted, "q"'] \arrow[d] & \{\sigma\} \arrow[r] \arrow[u, shift right, dotted, "q"'] \arrow[d] & \{\sigma\}/\{\pi\} \arrow[r] \arrow[u, shift right, dotted, "q"'] \arrow[d] & \{\beta\} \arrow[r] \arrow[u, shift right, dotted, "q"'] \arrow[d] & 0 \\
 & 0 & 0 & 0 & 0 & 0 &
\end{tikzcd}$$

In accordance with the notation of that article (TN 38, pp. 1-9), I have now used $(\Omega^A, \Pi_{A'})$, in place of $(\omega^A, \pi_{A'})$ as above, to denote constant twistors, therefore satisfying

$$\nabla_{AA'}\Omega_B = -i\varepsilon_{AB}\Pi_{A'}, \quad \nabla_{AA'}\Pi_{B'} = 0,$$

with, also,

$$\nabla_{AA'}\mathring{\Omega}_B = 0;$$

and the more general $(\omega^A, \pi_{A'})$ and $\mathring{\omega}^A$ now merely satisfy

$$\nabla_{A'}^{B}\omega_B = 2i\pi_{A'}, \quad \nabla_B^{A'}\pi_{A'} = 0$$

with

$$\nabla_{A'}^{B}\mathring{\omega}_B = 0.$$

The helicity 3/2 (Dirac-type) potential $\sigma_{A'B'C}$ (symmetric in A'B'), and the second potential $\rho_{A'BC}$ (symmetric in BC) satisfy

$$\nabla_{B'}^{B}\rho_{A'BC} = 2i\sigma_{A'B'C}, \quad \nabla_B^{B'}\sigma_{A'B'C} = 0$$

with

$$\nabla_{B'}^{B}\mathring{\rho}_{A'BC} = 0$$

(symmetric in BC) the gauge freedoms being

$$\rho_{A'BC} \mapsto \rho_{A'BC} - i\varepsilon_{BC}\pi_{A'} + \nabla_{CA'}\omega_B, \quad \sigma_{A'B'C} \mapsto \sigma_{A'B'C} + \nabla_{CB'}\pi_{A'}, \quad \mathring{\rho}_{A'BC} \mapsto \mathring{\rho}_{A'B'C} + \nabla_{CA'}\mathring{\omega}^B$$

The Frauendiener-Sparling-type quantities $\rho_{A'BC}$ and $\dot{\sigma}_{A'B'C}$ (still symmetric in BC and A'B', respectively) generalize these equations to

$$\nabla_{(B'}^{B}\dot{\rho}_{A')BC} = 2i\sigma_{A'B'C}, \quad \nabla_{(B}^{B'}\sigma_{C)A'B'} = 0, \quad \nabla_{(B'}^{B}\dot{\rho}_{A')BC} = 0$$

(with the same gauge freedom as before) and the Frauendiener-type quantities $\alpha_A, \beta_{A'}, \mathring{\alpha}_A$ arise as

$$\alpha_C = \tfrac{1}{2}\nabla^{BB'}\dot{\rho}_{B'BC}, \quad \beta_{A'} = \nabla^{BB'}\dot{\sigma}_{A'B'B}, \quad \mathring{\alpha}_A = \tfrac{1}{2}\nabla^{BB}\mathring{\dot{\rho}}_{B'BC}$$

and satisfy

$$\nabla_{A'}^{A}\alpha_A = 2i\beta_{A'}, \quad \nabla_A^{A'}\beta_{A'} = 0, \quad \nabla_{A'}^{A}\mathring{\alpha}_A = 0$$

All these relations hold consistently in Ricci-flat $\mathcal{M}$ — except for those involving Ω^A, $\Pi_{A'}$, and $\mathring{\Omega}^A$, which require $\mathcal{M}$ to be flat.

How are we to define the maps q, $\hat{q}$ in each of the various columns? In the case of

$$\begin{tikzcd} 0 \arrow[r] & \{\mathring{\Omega}^A\} \arrow[r, shift left, "i"] & \left\{\begin{matrix}\Omega^A\\ \Pi_{A'}\end{matrix}\right\} \arrow[l, shift left, dashed, "\hat{q}"] \arrow[r, shift left, "\rho"] & \{\Pi_{A'}\} \arrow[l, shift left, dashed, "q"] \arrow[r] & 0, \end{tikzcd}$$

these maps are just as defined earlier (for the flat case $\mathbb{T}^\alpha$), namely

$$\Pi_{A'} \overset{q}{\longmapsto} (iq^{AA'}\Pi_{A'}, \Pi_{A'}) \text{ and } (\Omega^A, \Pi_{A'}) \overset{\hat{q}}{\longmapsto} \Omega^A - iq^{AA'}\Pi_{A'}.$$

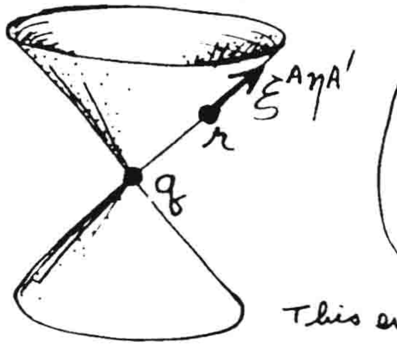
We need to generalize this to the remaining columns, and also in a way that makes sense in (Ricci-flat) $\mathcal{M}$. What this amounts to, in the case of the map q, is finding a way of *fixing* the "second potentials" ω_A, $\rho_{A'BC}$, $\dot{\rho}_{A'BC}$ (mod ω_A), and α_A, in terms of the "first potentials" and a given point $q \in \mathcal{M}$. This is achieved by requiring that the appropriate *null-data* for these second-potential quantities are *zero* on the light cone of q. In the case of the map $\hat{q}$, we need a way of obtaining, as fixed by the point $q \in \mathcal{M}$, a definite "free-field" quantity $\mathring{\omega}_A$, $\mathring{\rho}_{A'BC}$, $\mathring{\dot{\rho}}_{A'BC}$ (mod $\mathring{\omega}_A$), and $\mathring{\alpha}_A$, given the respective "sourced" quantity ω_A, $\rho_{A'BC}$, $\dot{\rho}_{A'BC}$ (mod ω_A), and α_A, with its respective "source" $\pi_{A'}$, $\sigma_{A'B'C}$, $\dot{\sigma}_{A'BC}$ (mod $\pi_{A'}$), and $\beta_{A'}$. This is achieved, in each case, by requiring that the null-data of the required "free fields" be equal to those of its corresponding "sourced fields" on the light cone of q.

It may be recalled (cf. Penrose & Rindler, Vol. 1, §§5.11, 5.12 ; R.P. (1980) Gen. Rel. Grav. 12, 225–64) that the null-datum at a point r on the data light cone, for a field $\theta_{A\ldots DP'\ldots S'}\ (=\theta_{(A\ldots D)(P'\ldots S')})$, is the quantity

$$\theta = \xi^A \ldots \xi^D \eta^{P'} \ldots \eta^{S'} \theta_{A\ldots DP'\ldots S'}$$

where $\xi^A\eta^{A'}$ is a (complex) tangent vector at r to the light cone. In order that the null-data for fields $\theta_{\ldots}, \ldots, \chi_{\ldots}$ determine, freely, the fields themselves, we require that these fields constitute an exact set (i.e. that their totally symmetrized n^{th} derivatives $(n=0,1,2,\ldots)$ be independent and sufficient to determine all the unsymmetrized m^{th} derivatives $(m=0,1,2,\ldots)$. In the present context, we find that the appropriate null-data are, respectively

$$\xi^A\omega_A\,,\ \begin{pmatrix}\eta^{A'}\xi^B\xi^C\rho_{A'BC}\\ \xi^A\xi^B\xi^C\nabla^{A'}_{A}\rho_{A'BC}\end{pmatrix},\ \begin{pmatrix}\eta^{A'}\xi^B\xi^C\dot{\rho}_{A'BC}\\ \xi^A\xi^B\xi^C\nabla^{A'}_{A}\dot{\rho}_{A'BC}\\ \xi^C\nabla^{BA'}\dot{\rho}_{A'BC}\end{pmatrix},\ \xi^A\alpha_A$$



except that $\eta^{A'}\xi^B\xi^C\dot{\rho}_{A'BC}$ is mere "gauge", and so does not contribute to $\{\dot{\rho}\}/\{\omega\}$

This enables all the q and $\hat{q}$ maps to be defined.

Thanks to Robin Graham for a valuable discussion in which he brought up the issue addressed by this article.

~ Roger Penrose

240

IN {2 2} VACUUMS, THE NULL-DATUM ON NULL RAYS IS EXACTLY r^{-5}

Gen. Relativity Gravitation **27** (1995), no. 12, 1323–5

See Chapter 223.

General Relativity and Gravitation, Vol. 27, No. 12, 1995

In $\{2\,2\}$ Vacuums, the Null-datum on Null Rays is Exactly r^{-5}

Roger Penrose[1]

Received August 21, 1995

It is shown, by a simple argument, that in any $\{2\,2\}$ (i.e. type D) vacuum space-time, the complex component of the Weyl curvature (spinor Ψ_{ABCD}) along any light ray is precisely proportional to r^{-5}, where r is the standard Schwarzschild radial coordinate, or its natural complex generalization for the arbitrary $\{2\,2\}$ case.

It has been known for some while [6] that any type $\{2\,2\}$ (i.e. type D) solution of the Einstein vacuum equations possesses a *Killing spinor* (see below). This holds whether or not there is a cosmological constant. The entire family of $\{2\,2\}$ solutions was exhibited by Kinnersley [4]. Included in this family are the familiar Schwarzschild, Kerr, NUT, and Levi-Civita C-metric solutions.

This Killing spinor can be taken to be

$$\kappa_{AB} = r\,\alpha_{(A}\beta_{B)}\,,$$

where the Weyl spinor for a type $\{2\,2\}$ vacuum has the form

$$\Psi_{ABCD} = 6K\,r^{-3}\alpha_{(A}\alpha_{B}\beta_{C}\beta_{D)}\,,$$

---

[1] Mathematical Institute, Oxford

(See Ref. 2 for all relevant notation and definitions.) In the case of the Schwarzschild solution, if we take the constant K to be given by $K = -mG$, where m is the mass and G is Newton's gravitational constant, then r is the usual Schwarzschild radial coordinate (i.e., for which the spheres $r =$ const., $t =$ const. have area $4\pi r^2$). For the Kerr solution, the quantity r is complex, and it provides a natural generalization of the Schwarzschild r-coordinate. In a general {2 2} solution, we simply *define* the coordinate r by the requirement that the above expression for Ψ_{ABCD} should hold. That is to say, we take

$$r = (\tfrac{3}{2}K^{-2}\Psi_{ABCD}\Psi^{ABCD})^{-1/6},$$

where K is some suitably chosen constant.

In fact, in the general case, the coordinate r is a "*complex luminosity parameter*" along each (double) principal null direction. That is to say, it satisfies the equation

$$Dr = -\rho r,$$

where D stands for propagation derivative along a principal null direction, and ρ is the complex convergence of the corresponding null rays. Taking the principal null direction to be that of α^A, we thus have

$$D = \alpha^A\bar{\alpha}^{A'}\nabla_{AA'}, \qquad \rho\alpha_A = \alpha^B\bar{\alpha}^{A'}\nabla_{AA'}\alpha_B.$$

(See Ref. 3, Fig. 7-2, on p.178, for an explanation of the meaning of ρ; it has an imaginary part which measures the rotation of the rays.) This property of r follows at once from the spin-coefficient expression for the equation satisfied by the Killing spinor κ_{AB}.

This equation is

$$\nabla_{A'(A}\kappa_{BC)} = 0,$$

where κ_{AB} is *symmetric*

$$\kappa_{AB} = \kappa_{BA}.$$

Now, let γ be any light ray (null geodesic) and choose the spinor ξ^A to be parallelly propagated along γ, with $\xi^A\bar{\xi}^{A'}$ being tangent to γ. It follows (see Refs. 6,3) from the Killing spinor equation on κ_{AB} that the quantity

$$\kappa = \kappa_{AB}\xi^A\xi^B$$

is *constant* along γ.

The correspondingly defined quantity

$$\Psi = \Psi_{ABCD}\xi^A\xi^B\xi^C\xi^D$$

is the complex component of the Weyl curvature directed along the ray (i.e. a quantity of the form $C_{abcd}\ell^a m^b \ell^c m^d$ in a standard null tetrad, with ℓ^a in the direction of γ). In fact Ψ is the *null datum* for the characteristic inital value problem on some null surface of which γ is the generator [1]. In the case of type {2 2}, we find from above

$$\Psi = 6Kr^{-3}(\alpha_A\,\xi^A\,\beta_B\,\xi^B)^2$$

and also

$$\kappa = r\,\alpha_A\,\xi^A\,\beta_B\,\xi^B$$

from which it follows immediately that

$$\Psi = 6K\kappa^2 r^{-5}$$

Thus, we have the following theorem.

Theorem. In any type {2 2} vacuum (with or without cosmological constant), the null-datum along any null ray is a constant multiple of r^{-5}.

It may be noted that the full parallelly propagated curvature Ψ_{ABCD} has no such simple dependence on r along γ. It is just that the particular component that is being "looked at" by γ at any one moment behaves in this remarkably simple way. (This result, first indicated in an unpublished Twistor Newsletter article by the author, features significantly in a computer-aided calculation carried out by Vanessa Thomas for the complete set of symmetrized derivatives of the curvature spinor at an arbitrary point in the Schwarzschild solution; see Ref. 5).

ACKNOWLEDGEMENTS

I am grateful to Alan Held for a valuable comment concerning the parameter r.

REFERENCES

1. Penrose, R. (1980) *Gen. Rel. Grav.* **12**, 225.
2. Penrose, R., and Rindler, W. (1984) *Spinors and Space-Time*, Vol. 1; *Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge).
3. Penrose, R., and Rindler, W. (1984) *Spinors and Space-Time*, Vol. 2; *Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge).
4. Kinnersley, W. (1969). *J. Math. Phys.* **10**, 1195.
5. Thomas, V. (1995). In *Further Advances in Twistor Theory. Volume II. Integral Systems, Conformal Geometry and Gravitaton*, L. J. Mason, L. P. Hughston, and P. Z. Kobak, eds. (Longman, Harlow, Essex).
6. Walker, M., and Penrose, R. (1970). *Commun. Math. Phys.* **18**, 265.

241

WHY NEW PHYSICS IS NEEDED TO UNDERSTAND THE MIND

Proc. Schrödinger Conference: What is life? The next fifty years: Speculations on the future of biology. Eds M. P. Murphy and L. A. J. O'Neill. Cambridge University Press, 1995

558 Why new physics is needed to understand the mind. *Proc. Schrödinger Conference: What is life? The next fifty years: Speculations on the future of biology.* Eds M. P. Murphy and L. A. J. O'Neill. 1995

9

Why new physics is needed to understand the mind

ROGER PENROSE
Mathematical Institute, Oxford

WHY CONSCIOUS UNDERSTANDING IS NON-COMPUTATIONAL

There are many facets to human mentality. It may well be that some of these can be explained in terms of our present-day physical concepts (compare Schrödinger, 1958) and, moreover, are potentially amenable to computational simulation. The proponents of artificial intelligence (AI) would maintain that such a simulation is indeed possible – at least for a good many of those mental qualities that are basically involved in our intelligence. Furthermore, such a simulation could be put to use in enabling a robot to behave, in those particular respects, in the same kind of way as a human being might behave. The proponents of *strong* AI would maintain, moreover, that *every* mental quality can be emulated – and will eventually be superseded – by the actions of electronic computers. They would also maintain that such mere computational action must evoke the same kind of conscious experiences in a computer, or robot, as we experience ourselves.

On the other hand, there are many who would argue to the contrary: that there are aspects of our mentality that cannot be addressed merely in terms of computation. Human consciousness, on such a view, would be such a quality – so it is *not* simply a manifestation of computation. Indeed, I shall argue so myself; but more than this, I shall argue that those actions which our brains perform in accordance with conscious deliberations must be things that cannot even be *simulated* computationally – so certainly computation cannot of itself give rise to any kind of conscious experience.

In order that such arguments can be made precise, it is necessary to have a clear-cut notion of what is *meant* by a 'computation'. Indeed, there is a mathematically precise definition of computation. This can be given in terms of what are known as *Turing machine* actions. A Turing machine is a mathematically idealized computer – idealized so that it can run on indefinitely without wearing out or slowing down, so that it never makes mistakes, and, most importantly, so that it has an unlimited storage capacity. (Thus, we must imagine that there is always the possibility for more storage capacity to be added, if there were danger of it running out.) I do not propose to give a more precise definition of a Turing machine than this, since the notion of 'computer' is now somewhat familiar to us. (For more details, see, for example, Penrose, 1989.)

As for the notion of 'consciousness', I shall make no attempt to define it here. All that we shall need about this notion is that whatever consciousness is, it is something that is necessarily present when we *understand* – when, in particular, we understand a mathematical argument.

Why do I claim that the effects of conscious deliberation cannot even be *simulated* by computational procedures? My own reasons stem, most powerfully, from the famous theorem of Kurt Gödel (1931). Gödel's theorem has the clear implication that mathematical understanding cannot be reduced to a set of known and fully believed computational rules. It is possible to go further than this and argue that no knowable set of purely computational procedures could lead to a computer-controlled robot that possesses genuine mathematical understanding. Such procedures could include not only deliberate 'top-down' algorithmic instructions, but also some more loosely programmed 'bottom-up' learning mechanisms. It is not really appropriate for me to enter into details; the full arguments will be given elsewhere (Penrose, 1994).

It would be unreasonable to suppose that there is anything particularly special about mathematical understanding, as opposed to other kinds of human understanding, with regard to the issue of non-computability. Accordingly, the non-computability of our mathematical understanding presumably has the implication that *every* kind of human understanding is also achieved by non-computable means. Likewise, it seems to me to be unreasonable to suppose that the various other aspects of human consciousness can be explained computationally any more than understanding can. Finally, I believe that non-human animals – at least many different kinds of animals – also possess the quality of consciousness and, consequently, must also act according to non-computational rules.

THE TWO LEVELS OF PHYSICAL ACTION

For the purpose of the remainder of our discussion, let us indeed take it that our brains act non-computationally when we indulge in conscious thought processes. Let us also accept that the actions of our brains are entirely governed by those same physical laws that underlie the behaviour of inanimate matter. Then we are confronted by the requirement that there must be physical actions that are governed by physical laws, yet which *in principle* cannot be simulated entirely computationally. What actions might these be?

We must first try to see whether there is scope, within the physical laws that are presently understood, for an appropriate non-computational behaviour. If we find that these laws fail to offer the scope that we need, then we must look beyond these laws in order to find the necessary non-computational processes. We must also ask for a plausible place where such non-computational physics could provide a key input into the functioning of our brains.

What, then, is the picture that today's physicists provide for us – as to the precise way that the physical world is understood to act? They would maintain that at the most fundamental level, the laws of quantum mechanics must hold. According to the Schrödinger picture, the state of the world at any one moment would be described by a *quantum state* (often denoted by the letter ψ, or by $|\psi\rangle$ in Dirac's notation) which describes a weighted combination of all the alternative ways in which the system under consideration *might* behave. It is not a probability-weighted combination, because the weighting factors are *complex numbers* (i.e. numbers of the form $a+ib$, where $i^2=-1$, a and b being ordinary real numbers). Moreover, the time-evolution of the quantum state is governed by a clear-cut deterministic equation, called the *Schrödinger equation.* The Schrödinger equation is a *linear* equation (leaving these complex weighting factors unaltered) and, in any ordinary sense, it would certainly be considered as providing a *computable* evolution for the quantum state. Thus quantum theory, in this sense, does not give us anything essentially non-computable.

However, Schrödinger evolution alone does not provide a picture of the world that makes sense at the classical level of phenomena (as Schrödinger himself was careful to emphasize). The rules of quantum linear superposition seem to apply only to states that differ very little from each other. Two states that differ very greatly from each other – such as two discernibly different locations for a golf ball – do not appear to exist in linear superposition. A golf ball, for example, has one location or it has another. It does not find itself in two places at once. An electron or a neutron, on the other hand, *can* exist in a superposition of two completely different locations at once (with complex-number weighting factors), and many experiments have been performed to confirm this kind of thing.

Thus, it appears, we must consider that there are two different levels of physical phenomena. There is the 'small-scale' *quantum* level, at which particles, atoms, or even molecules, can exist in these strange complex-number-weighted quantum superpositions. Yet, on the other hand, there is the *classical* level, at which one thing can happen, or another thing can happen, but where we do not experience complex combinations of alternatives. In particular, a golf ball is a classical-level object.

Of course, classical-level objects such as golf balls are themselves built from quantum-level constituents like electrons and protons. How is it that there can be one set of rules for the constituents and another for the large-scale object itself? In fact, this is a delicate issue which is not fully resolved within today's physical pictures. I shall need to return to the matter shortly; but, for the moment, it is best that we simply take the view that there are indeed two distinct levels of physical behaviour, and that there are different laws governing each.

HOW ARE THESE LEVELS BRIDGED?

At the quantum level, the mathematical description of a physical system is provided by the quantum state $|\psi\rangle$, as referred to above – sometimes called the system's *wavefunction*. So long as the system remains at the quantum level this state evolves with time (in the Schrödinger picture), in accordance with the deterministic and computable Schrödinger equation. I shall denote this evolution by **U** (unitary evolution). At the entirely classical level, the laws that control physical objects are those of Newton (for the commonplace motions of ordinary objects), of Maxwell (for the behaviour of electromagnetic fields), and of Einstein (when velocities or gravitational potentials become large). For all these classical types of evolution I shall use the notion **C**. These laws are again of a deterministic nature, and seem to be basically *computable* also. (In my assertions that **U** and **C** are 'computable', I have glossed over the fact that both **U** and **C** operate with continuous rather than the discrete parameters that are relevant to Turing-computability. We may suppose that adequate discrete approximations to **U** and **C** can be employed for this purpose, although this is less clear cut in the case of **C**, which is frequently chaotic, than it is in the case of the linear **U**.)

But how, in standard physical theory, does one deal with processes that involve both levels at once? Suppose, for example, that a physical system is so delicately poised that the behaviour of one of its quantum-level constituents can trigger off a large-scale classical effect? This is just the kind of situation that is referred to, in quantum theory, as a 'quantum measurement', and a different type of description from that entailed by Schrödinger's equation is needed. This is referred to as *state-vector reduction* (or collapse of the wavefunction), and I shall denote it by **R**. The standard mathematical procedure for describing a 'measurement' in quantum mechanics involves an instantaneous 'jump' from one quantum state to another. It is in the 'jumping' that is involved in quantum theory's **R**-procedure that all the probabilities and uncertainties of evolution arise; for a system that remains entirely at the quantum level, its evolution is described by the entirely deterministic and computable procedure **U**.

In any particular quantum measurement, the different possible outcomes are determined by the specific nature of the measurement that is being performed. All that the theory tells us is that there are certain possibilities attached to these various outcomes, these probabilities being determined by the particular quantum state that is being subjected to measurement. The theory makes no definite statement as to *which* of the possible outcomes takes place (except in special situations where the probabilities are 1 or 0). With regard to the results of measurements, *probability assignments* are all that the theory comes up with. Within the limitations that are provided by these probabilities the behaviour of the system is entirely *random*, whenever a measurement is performed.

Thus, present-day physical theory tells us that the objects of this world behave, for the most part, in an entirely computable way, except that from time to time (i.e. when a 'measurement' – or something equivalent – takes place) there is an additional entirely random ingredient in the behaviour of the system. Without this random ingredient, any physical system's behaviour would be regarded as *computational*, in the sense that a Turing-machine simulation could be set up which approximates the behaviour of the system as closely as is desired. Thus, we may regard a general physical system – as far as our present-day descriptions go – as something which behaves as a Turing machine with a randomizer.

However, a 'randomizer', in this sense, does not *in effect* give us anything that is beyond ordinary Turing computability. In practice, effectively random behaviour can be achieved by means of what are called 'pseudo-random' procedures, which means computational procedures that behave as random ones, for all intents and purposes. What would normally be done would be to have some 'chaotic calculation' which, though entirely computational in nature, would depend very critically upon some starting parameter. One might choose, for example, the precise *time* (as measured by the computer's clock) for this starting parameter. The result of this calculation would be, in effect, completely random, although it would be the result of the action of a Turing machine. There would be no difference in practice between using a pseudo-random computation of this nature and a genuinely random one.

The picture of physical reality provided by something that could be accurately modelled by a Turing machine with a randomizer – or with a pseudo-random input – does not, however, provide us with the kind of *non-computability* that the 'Gödelian' arguments of Section 1 tell us are necessary for the actions of a conscious brain. But is the 'pure randomness' that standard theory gives us *really* what is going on in a physical system? The weakest point in our present-day physical picture, at least at the kind of level that might be relevant to brain action, indeed lies in the random **R** process. Perhaps the present-day use of an entirely random '**R**' is merely a stop-gap? In my own opinion, this is indeed the case, and some new physical insights and a *new* physical theory will be needed in order to bridge the gap between **U** and **C**. In fact there is an increasing acceptance among (still a minority of) physicists that something needs to be done.

THE GRW STATE-REDUCTION SCHEME

One of the most promising recent suggestions for a modification of the rules of quantum theory, in accordance with aspirations of this nature, is that put forward by Giancarlo Ghirardi, Alberto Rimini and Tullio Weber (GRW). In their original scheme (Ghirardi, Rimini & Weber, 1986), GRW proposed that although a particle's wavefunction would evolve precisely according to the Schrödinger equation **U** for *most* of the time, there would be a small probability that the wavefunction would suffer a 'hit', whereby the wavefunction would be multiplied by another function, with a Gaussian spatial dependence. There are two arbitrarily chosen parameters in this theory, one of which (call it λ) determines the width of the Gaussian function, and the other (call it τ) which determines the rate at which the hits are taken to occur. The location of the peak of the Gaussian function is taken to be random, but with a probability distribution governed by the squared modulus of the wavefunction at the time at which it suffers the 'hit'. In this way, an agreement is achieved with the standard 'squared-modulus rule' governing the probabilities of ordinary quantum theory.

In the original GRW proposal, the value of τ is chosen so that a single particle on its own would suffer a 'hit' about once every 10^8 years. Thus, there is no conflict, over ordinary periods of time, with the standard quantum-mechanical descriptions of individual particles (whence, for example, the neutron diffraction experiments of Zeilinger *et al.*, 1988 are consistent with GRW). However, for systems involving large numbers of particles, the phenomenon of *quantum entanglement* must be taken into account. I shall describe this important phenomenon shortly, but for the moment, we just take note of the fact that, in standard quantum theory, the wavefunction of a system involving many particles must refer to the system in its entirety, and there is not simply a separate wavefunction for each particle individually. Thus, for a classical-level object with a large number of particles (say a golf ball), as soon as *one* of the constituent particles suffers a hit, the *entire* wavefunction of the object would be reduced. In the case of a golf ball, for which the number of particles would be about 10^{25}, the state would be reduced in less than a nanosecond. Thus, a quantum state which consists of a superposition of a golf ball in one location and the same golf ball in another location would be reduced, in a timescale of less than a nanosecond, *either* into the quantum state in which the golf ball has one of these locations *or* into the quantum state in which it has the other location.

In this way, the GRW scheme resolves one of the most fundamental problems encountered by the standard version of quantum theory: the *paradox of Schrödinger's cat* (Schrödinger, 1935a). According to this paradox, a cat is placed in a quantum superposition of two states, in one of which the cat is alive, and in the other, it is dead. Standard quantum theory – taken to insist that the evolution of a quantum state takes place solely according to the U-process – would tell us that the superposition of dead and alive cats must persist, and cannot resolve itself into one or the other. However, in the GRW scheme, the cat's state would indeed resolve itself into one or the other – in a timescale of a good deal less than a nanosecond.

ENTANGLED STATES

An important feature of the above scheme is that it depends upon the fact that a quantum state involving many particles is likely to be what is called an *entangled state*. I shall illustrate this type of situation in terms of what are called EPR (Einstein–Podolsky–Rosen) phenomena. These also serve to emphasize the essentially *non-local* nature of quantum states in relation to the **R**-process.

Imagine an initial state of spin 0 which decays into two particles, each of spin 1/2, these travelling off in opposite directions. If some specific direction in space is chosen, and the spin of each of these two particles is measured in that direction, then it has to be the case that the *opposite* answer is obtained for each particle – because the combined state of spin is zero. This applies whatever particular direction happens to be chosen.

More complicated measurements can also be performed, in which a *different* spin direction is chosen for each of the two particles. The measurement, in each case, yields merely the answer 'yes' or 'no' (since a spin 1/2 particle carries just one bit of information with regard to its spin). But there are certain joint probabilities, determined by standard quantum theory, for the answers for the two particles to *agree*, or to *disagree* (in fact $1-\cos\theta:1+\cos\theta$, where θ is the angle between the angles chosen).

Now according to a famous theorem of John S. Bell (1964), there is actually no 'local' way of explaining the joint probabilities that describe the quantum-mechanical predictions for these pairs of measurements, where each particle is regarded as a separate entity on its own. One must consider that the two particles are, in some mysterious way, still 'connected' to each other right up until a measurement is made on one or other of the particles. In effect, we have the situation that as soon as such a measurement on *one* particle is made, then this measurement instantaneously causes the state of the *other* particle to be reduced also. The state of the pair of particles cannot be thought of as being given by one particular state for one particle together with another particular state for the other particle. The *pair* of particles has a quantum state – an *entangled* state – but neither particle, separately, has a state on its own.

The phenomenon of quantum entanglement was first described by Erwin Schrödinger (1935b), as a general feature of quantum systems. Bell's theorem opened the way to experimental verification of the effects of quantum entanglements over large distances. These effects were subsequently observed by a number of experimenters, the most impressive experiments of this nature being performed by Alain Aspect and his colleagues (1982), where entanglements over distances of some twelve metres were observed.

Any ordinary large-scale object, such as the cat of Schrödinger's thought experiment, as described above, would itself be an entangled system. The individual particles of the cat's body would not have states on their own, but would be part of an entangled state for the cat as a whole. This would indeed be an implication of the standard quantum-mechanical descriptions. Now, the GRW scheme uses just the same kind of quantum-mechanical descriptions. Quantum entanglements are indeed just as much a feature of that scheme as they are for standard quantum mechanics. In this way, as soon as one of the particles of Schrödinger's cat's body suffers a GRW 'hit', the state of the entire cat is reduced with it, so the cat's state becomes either 'dead' or 'alive', rather than a quantum superposition of both together.

ENTANGLEMENT WITH THE ENVIRONMENT

In fact, the state of the cat would not be isolated from that of its environment, and we would have to consider that the entanglements do not end with the cat, but extend into this environment as well. Moreover, there would be many more particles involved in this disturbed environment than there would in the cat's body itself. In the *standard* discussions of the measurement process, the role of the environment of a quantum system is indeed considered to be all important. As the standard argument goes, the detailed quantum information (in what are called 'phase relations') that distinguishes a quantum superposition from a probability-weighted combination will simply get lost in these entanglements with the environment. Thus, *in effect* – or what John Bell has referred to as FAPP ('for all practical purposes') – a quantum superposition will behave like a probability-weighted combination of alternatives, as soon as entanglements with the random environment become significant.

However, all that this standard argument really achieves is a *coexistence* (FAPP) between the quantum procedures **U** and **R**, rather than a *deduction* of **R** from **U** (which would, in any case be impossible, strictly speaking, if only because the **U**-procedure itself makes no mention of probabilities). We still need something more than merely deterministic Schrödinger evolution

U of a system, if we are to explain how physical objects actually behave (as Schrödinger himself was careful to emphasize). What schemes like that of GRW attempt to achieve is a picture in which the physically observed process **R** – or something very like it – becomes part of the actual physical evolution of a system.

In fact, in the GRW scheme, it would be in the environment of a system that the 'hits' would normally occur first, and the entanglements of this environment with the system would result in the reduction procedure **R** taking effect in the system itself. For example, a DNA molecule would be far too small for 'hits' to be significant in the individual nucleotides of the molecule itself. Without 'hits' in the entangled environment playing their additional role, there would be nothing to determine the DNA strand as having a particular succession of nucleotides rather than merely having a quantum superposition of a number of different ones!

A NON-COMPUTABLE GRAVITATIONAL REDUCTION SCHEME?

None of this says anything about a role for non-computability in physical action, whose necessity was strongly argued for in Section 1. I have so far merely pointed to an important gap in our physical understanding – namely at the quantum/classical-level borderline – and mentioned one particular proposal (the GRW scheme) which attempts to bridge this gap. I would argue that there are strong reasons to believe that the physics which must *actually* bridge this gap will result from the appropriate union between quantum theory and Einstein's general theory of relativity. It is generally accepted that this union would have to result in a change in Einstein's gravitational theory (at very tiny distance scales), but it is a less conventional viewpoint that the standard rules of (U-)quantum theory must also have to change when the appropriate union is found, so that the phenomenon of state-vector reduction **R** will turn out to be a quantum-gravitational phenomenon (see Komar, 1969; Károlyházy, 1974; Károlyházy *et al.*, 1986; Diósi, 1989; Ghirardi *et al.*, 1990; Penrose, 1989, 1993, 1994).

If one accepts that the missing theory that is needed in order to replace the stop-gap **R** must be a gravitational one, then one is led to certain 'order-of-magnitude' estimates of the levels and timescales at which **R** ought actually to take place. (There are also some indirect – and rather tentative –

indications that such a theory might well be of a non-computational nature; cf. Penrose, 1994.)

To get some understanding of the level at which such a theory ought to begin to have relevance, consider the situation where a solid lump of material is placed in a quantum linear superposition of two distinct locations. I shall suppose that such a superposition is like an unstable particle or nucleus, with a certain half-life, and two separate modes of decay. In one of these modes, the superposed state decays into that state in which the lump occupies one of the two locations under consideration, and in the other mode, it decays into the state in which the lump occupies the other location. To estimate the half-life of this decay, we consider the energy E that it would take to displace the two instances of the lump away from each other, starting from coincidence, to the separation that they have in the superposition under consideration, where we take into account only the effect that the *gravitational* field of one of the lumps would have on the other (compare also Diósi, 1989). Another way of saying the same thing, assuming that the lump moves rigidly, is that E is the gravitational self-energy of the *difference* between the Newtonian gravitational fields of the two instances of the lump (cf. Penrose, 1994). The half-life T for the decay of the superposed state into one state of location or the other is then of the order of

$$T = \hbar/E,$$

where $\hbar$ is Planck's constant divided by 2π.

Let us examine this criterion in certain simple situations. If the lump consisted of just a single nuclear particle, then (taking the particle's radius to be of the order of one fermi) we get a decay time of some 10^7 years – similar to that of the original GRW scheme. For a lump of water density, we find that if its radius were one micrometre, then T would be about one twentieth of a second. If its radius were 10^{-3}cm, then T would be less than a millionth of a second; for a radius of 10^{-5}cm, then a few hours.

However, as argued above, these would be the reduction times only if the lump could remain isolated from its environment. If significant amounts of environmental material become disturbed, then the reduction time could be much shorter.

The proposal here is that for this state-reduction scheme to exhibit significant non-computational features, it would be necessary that the reduction takes place in the *system itself* – I shall refer to this as *self-reduction* – rather than in the environment. The idea is that the environment is essentially random, so that any genuinely non-computational features would be masked by this randomness, so long as it is state-reduction in the system's environment that (because of entanglement effects) reduces the system's state. In any normal experimental situation, it would indeed be the environment that controls the reduction of the state, so we see nothing different from the normal random behaviour that is so effectively described by the standard quantum-mechanical procedure **R**. It would take a very carefully organized structure for sufficient quantum isolation to occur that self-reduction takes place – before the random environment takes over. Such organization would be required so that there are significant *non-computational* deviations from the normal random **R**-procedure, as would be required according to the considerations of Section 1. No physical experimental set-up has yet got close to obtaining the necessary isolation.

RELEVANCE TO CONSCIOUS BRAIN ACTION

This is not to say that Nature herself has found no way to achieve the necessary conditions. Indeed, the arguments of Section 1 must be telling us that somehow she *has* found such a way. The conventional picture of brain action is that it can be completely understood in terms of nerve signals and synaptic action. Nerve signals seem to disturb their environment far too much for anything close to the isolation needed for the criteria of Section 7 to be achievable. But what about synaptic action? The strengths of (at least some) synapses are subject to continual change. What is it that controls these changes? There seem to be various different possibilities and different proposals, but one important factor seems to be the activity of the *microtubules* in the cytoskeletons of neurons.

What are microtubules? They are tiny tubelike structures which inhabit eukaryotic cells generally and which play many different kinds of role within cells. For example, in a one-celled animal they seem to be important in controlling its locomotion – as is the case for the continual change of shape of an amoeba. They exist in individual neurons, and control the ('amoeba-like') way that they connect with other neurons. They extend along the lengths of axons (perhaps not all in one piece) and dendrites, in each case right up to the close vicinities of the synapses. They transport various molecules along their lengths – in particular, the neurotransmitter chemicals that are vital for propagating neuronal nerve signals across synapses.

570 Why new physics is needed to understand the mind. *Proc. Schrödinger Conference: What is life? The next fifty years: Speculations on the future of biology.* Eds M. P. Murphy and L. A. J. O'Neill. 1995

Mictrotubules are composed of a peanut-shaped protein called *tubulin* (about 8 nm × 4 nm × 4 nm in dimension), where the tubulins are arranged in a slightly skew hexagonal lattice. Each tubulin is a 'dimer', consisting of two components called 'α-tubulin' and 'β-tubulin'. A tubulin dimer is capable of existing in (at least) two different states – called 'conformations' (apparently depending upon the location of an electron, centrally placed in a 'hydrophobic pocket' between the two components). It has been suggested by Hameroff and Watt (1982; cf. also Hameroff, 1987) that these conformations give individual microtubules computer-like properties, where the two conformations of the dimers behave like the states 'on' or 'off', coding the bits '1' or '0' in a computer. Complicated signals could propagate along the microtubules in the manner of a cellular automaton.

So far this merely gives the potentiality for *computer*-type actions with enormously greater potential power than would be the case if we take individual neurons to be the sole 'computational units'. (Tubulin conformations act about a million times faster than neuron signals, and there are about ten million tubulins per neuron.) However, according to the foregoing discussion, we need something more than this, namely some scope for a *non-computational* action – an action which could arise only if a large-scale quantum-coherent state can be maintained in an isolated environment sufficiently long that the quantum state (or, at least, parts of the quantum state) can *self-collapse*, rather than collapsing because of entanglement with the environment. Do microtubules provide a plausible place for this kind of physical action? I believe that there is a good prospect that this may be so. Recall that microtubules are *tubes*. There is scope for some kind of quantum oscillation taking place *within* the tubes (Hameroff, 1974; del Giudice *et al.*, 1983; Hameroff, 1987; Jibu *et al.*, 1994; cf. also Fröhlich, 1968), which could be weakly coupled to the conformational actions, taking place along the tubes, of the tubulin dimers. The quantum oscillations within the tubes would probably not involve a significant amount of mass movement, but situations could arise when the coupling with the tubulin conformations becomes sufficiently great that there is just enough mass movement for self-collapse to occur. The viewpoint being presented here is that there is a non-computability involved in this self-collapse, and that conscious events are in some way to be identified with this process.

Clearly there is a good deal of speculation involved in these suggestions, but it seems to me that *something* of this general nature is needed. A much more complete presentation of these arguments is to be found in Penrose (1994) and a further article in preparation by Hameroff and Penrose.

Why new physics is needed to understand the mind. *Proc. Schrödinger Conference: What is life? The next fifty years: Speculations on the future of biology.* Eds M. P. Murphy and L. A. J. O'Neill. 1995

REFERENCES

Aspect, A., Grangier, P. & Roger, G. (1982). Experimental realization of Einstein–Podolsky–Rosen–Bohm *Gedankenexperiment*: a new violation of Bell's inequalities. *Physical Review Letters* **48**, 91–4.

Bell, J. S. (1964). On the Einstein–Podolsky–Rosen paradox. *Physics* **1**, 195–200. Reprinted (1983) in *Quantum Theory and Measurement*, eds. J. A. Wheeler and W. H. Zurek Princeton: Princeton University Press.

del Giudice, E., Doglia, S. & Milani, M. (1983). Self-focusing and ponderomotive forces of coherent electric waves – a mechanism for cytoskeleton formation and dynamics. In *Coherent Excitations in Biological Systems*, eds. H. Fröhlich & F. Kremer. Berlin: Springer.

Diósi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Physical Review A* **40**, 1165–1174.

Fröhlich, H. (1968). Long-range coherence and energy storage in biological systems. *International Journal of Quantum Chemistry* **II**, 641–649.

Ghirardi, G. C., Rimini, A. & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review* **34**, 470.

Ghirardi, G. C., Grassi, R. & Rimini, A. (1990). Continuous-spontaneous-reduction model involving gravity. *Physical Review A* **42**, 1057–1064.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik* **38**, 173–198.

Hameroff, S. R. (1974). Chi: a neural hologram? *American Journal of Clinical Medicine* **2**(2), 163–170.

Hameroff, S. R. (1987) *Ultimate Computing. Biomolecular Consciousness and Nano-Technology*. Amsterdam: North Holland.

Hameroff, S. R. & Watt, R. C. (1982). Information processing in microtubules. *Journal of Theoretical Biology* **98**, 549–61.

Jibu, M., Hagan, S., Hameroff, S. R., Pribram, K. H. & Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *BioSystems* **32**, 195–209.

Károlyházy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat* **12**, 24.

Károlyházy, F., Frenkel, A. & Lukács, B. (1986). On the possible role of gravity on the reduction of the wave function. In *Quantum Concepts in Space and Time*, eds. R. Penrose & C. J. Isham. Oxford: Oxford University Press.

Komar, A. B. (1969). Qualitative features of quantized gravitation. *International Journal of Theoretical Physics* **2**, 157–160.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Penrose, R. (1993). Gravity and quantum mechanics, in *General Relativity and Gravitation 1992. Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina, 28 June–4 July 1992. Part 1: Plenary Lectures*, eds. R. J. Gleiser, C. N. Kozameh & O. M. Moreschi. Bristol: Institute of Physics Publishing.

Penrose, R. (1994). *Shadows of the Mind: An Approach to the Missing Science of Consciousness*. Oxford: Oxford University Press.

Schrödinger, E. (1935a). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften* **23**, 807–812, 823–828, 844–849. (Translation by J. T. Trimmer (1980). *Proceedings of the American Philosophical Society* **124**, 323–38. Reprinted in *Quantum Theory and Measurement*, eds. J. A. Wheeler & W. H. Zurek. Princeton: Princeton University Press, 1983.)

Schrödinger, E. (1935b). Probability relations between separated systems. *Proceedings of the Cambridge Philosophical Society* **31**, 555–563.

Schrödinger, E. (1958). *Mind and Matter.* Reprinted (1967) in *What is Life?* with *Mind and Matter* and *Autobiographical Sketches*. Cambridge: Cambridge University Press.

Zeilinger, A., Gaehler, R., Schull, C. G. & Mampe, W. (1988). Single and double slit diffraction of neutrons. *Reviews in Modern Physics* **60**, 1067.

242

THE ROLE OF GRAVITY IN QUANTUM STATE REDUCTION

Proc. Int. Conf. Non-accelerator particle physics (1994). Ed. R. Cowsik. World Scientific, Singapore 1995. Also published in *Astrophysics and Cosmology.* Eds B. Sinha and R. K. Moitra, Narosa Publishing House, 1995

574 The role of gravity in quantum state reduction. *Proc. Int. Conf. Non-accelerator particle physics* (1994). Ed. R. Cowsik. 1995. Also published in *Astrophysics and Cosmology.* Eds B. Sinha and R. K. Moitra, 1995

THE ROLE OF GRAVITY IN QUANTUM STATE REDUCTION

by ROGER PENROSE

Mathematical Institute
Oxford, U.K.

1. *General Relativity and Quantum Mechanics*

As was pointed out by Eugene Wigner in a famous lecture [1], mathematical theories have an extraordinary effectiveness in modelling the behaviour of the physical world in some of its most fundamental respects. We have become accustomed to the immensely impressive eleven or twelve figure accuracy of quantum electrodynamics, but more recently this precision has been overtaken by general relativity, as the binary pulsar PSR 1913+16 observations of Taylor and Hulse have demonstrated [2]. Taken over a twenty year period, the precision in the timing of this system is accurate to something like one part in 10^{14}, and it agrees with the detailed picture provided by general relativity – from the Newtonian aspects of the orbits, at one end of the scale, to the minute corrections from the energy loss due to gravitational radiation, at the other – to this same precision. This remarkable achievement has now been officially acclaimed by the award of the 1993 Nobel Prize in physics.

Despite the extraordinary accuracy of both quantum (field) theory and general relativity, each has its fundamental problems. In the case of general relativity, the basic inherent difficulty is the presence of space-time singularities in generic solutions of the Einstein equations, with an energy-tensor satisfying the normal inequalities for classical matter. These singularities are virtually unavoidable for space-times which qualitatively resemble the standard Friedmann-Lemaitre-Robertson-Walker (FLRW) cosmological models or the standard Oppenheimer-Snyder gravitational collapse model, as the classical singularity theorems [3] show.

Thus, classical general relativity does not allow us to avoid the problem of the big-bang singularity or of the singularities in black holes. The normal attitude to these fundamental problems is that they will be resolved as soon as the procedures of quantum theory can be successfully brought to bear on general relativity - just as the singularities in the classical Rutherford atom (arising from the classical spiralling of electrons into the nucleus) are resolved when quantum theory procedures are correctly applied to it. Accordingly, so it would be argued, the correctly quantized theory of gravity should also be singularity-free in the appropriate sense.

What are the fundamental problems that confront quantum theory? In the first place there are the divergent integrals and divergent sums that are ever-present in perturbative quantum field theory. It is an accepted expression of hope that when gravity is correctly incorporated into quantum field theory then this will supply the needed short-distance cut-off (at $\sim 10^{-33}$cm.); accordingly the overall theory may become finite. Secondly, the measurement paradox continues to provide a profound obstruction to our acceptance of a coherent quantum-mechanical picture of physical reality. I shall be presenting, arguments in favour of my own belief that the measurement paradox will itself not be resolved until the appropriate union between quantum theory and general relativity is found. By such a union I do not mean a "quantum gravity" theory of the type that workers in the field normally aspire to - that is to say, a theory obtained by applying the rules of standard quantum theory to general relativity (or else to some other theory of gravity if quantizing Einstein's particular theory turns out to be inappropriate). I believe, instead, that there will be as much change in the very structure of quantum mechanics as there will be in the space-time of Einstein's general relativity when the correct union of the two is eventually found. As we have seen, there is no less accuracy in Einstein's classical theory than there is in quantum theory itself.

2. *Time-asymmetry in Space-Time Singularities*

One of the main indications that the correct union of general relativity with quantum mechanics may involve a change in the quantum theory lies in the structure of space-time singularities. There is good observational evidence for the big-bang singularity being of an extremely uniform isotropic nature, whereas it is to be expected that the singularities that arise in gravitational collapse must be highly non-isotropic and irregular [4]. This is intimately connected with the second law of thermodynamics. According to the second law, the initial state of the universe must have been one of low entropy, whereas the entropy would be very high in the final state. The low state of entropy in the initial state had nothing to do with its matter content; for a state of thermal equilibrium with regard to the matter is an assumption that is eminently consistent with all the successful predictions of the standard hot-big-bang model. Instead, the special nature of this low-entropy initial state lies in its very particular space-time geometry. The enormous numbers of available gravitational degrees of freedom were not even called upon in this initial state, and owing to the extreme weakness of the gravitational coupling it has taken - and is still taking - a very long time for these gravitational degrees of freedom to become "thermalized".

To understand how these gravitational degrees of freedom are manifested in space-time geometry, it will be helpful to examine how the full Riemann curvature tenson of space-time splits into two (roughly equal) portions : the Weyl conformal tensor and (essentially) the Ricci tensor,

$$R_{abcd} = C_{abcd} + E_{abcd}$$

where the Weyl tensor is defined by [5]

$$C_{ab}{}^{cd} = R_{ab}{}^{cd} - 2R_{[a}{}^{[c}\, g_{b]}{}^{d]} + \frac{1}{3} R\, g_{[a}{}^{c}\, g_{b]}{}^{d}$$

(square brackets denoting anti-symmetrization), and where

$$E_{ab}{}^{cd} = 2R_{[a}{}^{[c}\, g_{b]}{}^{d]} - \frac{1}{3} R\, g_{[a}{}^{c}\, g_{d]}{}^{d}$$

is determined by the Ricci tensor.

Einstein's equations tell us that the Ricci tensor is directly specified in terms of the energy tensor T_{ab} of matter (G being Newton's gravitational constant). Thus R_{ab}, or equivalently E_{abcd}, acts directly in providing the "source" for the gravitational field, in analogy with the charge-current vector J_a of Maxwell's electromagnetic theory. Analogous to the Maxwell field tensor (though at a different order of differentiation) is Weyl's tensor C_{abcd}. In this way, the pure gravitational degrees of freedom may be thought of as being described by C_{abcd}. Thus, the fact that the gravitational degrees of freedom were not made use of in the big-bang is expressed in the vanishing of the Weyl curvature at this initial singularity.

Indeed, the standard FLRW cosmological models do have vanishing Weyl curvature, and they appear to resemble the structure of our own universe very closely in its initial singular state. On the other hand, the singularities in black holes - and in what would be expected in a generic all-embracing big crunch - are of a completely different character. Here, the Weyl curvature dominates, and diverges to infinity as the singularity is approached. It appears that "quantum gravity" has very different things to say about past (big-bang) and future (black holes and big crunch) singularities. Apparently it says something of the nature of the *Weyl curvature hypothesis*, which asserts. "The Weyl curvature tends to zero as a space-time singularity is approached from the future time-directions".

3. *Geometrical Interpretation of Weyl and Ricci Tensors*

It is instructive to examine the different effects of the Weyl and Ricci tensors on a system of freely falling test particles. Imagine an observer (a test particle) in free motion under gravity, in empty space - say in orbit about the earth. Suppose that the observer is surrounded by a small sphere of other test particles, held initially at rest with respect to the observer. Owing to the tidal effect of the earth's gravitational field, the sphere becomes distorted into an ellipsoid as the test particles each follow their slightly differing orbits. This tidal distortion, being entirely due to the vacuum's *Weyl* curvature, is such that the *volume* of the ellipsoid into which the sphere is initially distorted is *equal* to that of the sphere.

On the other hand, if the sphere were to surround a region containing matter, then there would be an initial volume-reducing effect which is roughly proportional to the total mass enclosed – though there is an additional contribution due to the pressures. This volume reduction is an effect of the *Ricci* curvature. It is most clearly illustrated in the situation where the sphere surrounds the entire earth. The acceleration due to gravity is inwards at all points of the sphere, and the volume is indeed reduced. Note that the equal-volume distortion due to Weyl curvature and the volume reduction due to Ricci curvature are both to be taken at the level of initial *accelerations*, where the test particles are taken to be at initially at rest with respect to the central point.

The division into distortion and reduction effects is somewhat cleaner if we think in terms of light rays [6] rather than massive test particles. Here, the effect of Ricci curvature is to provide a focussing of the light rays that is isotropic in the spacelike directions perpendicular to the direction of the rays. That is to say, Ricci curvature acts as an anastigmatic lens. The lens is positively focussing if the energy component in the (null) direction along the ray is positive. In contrast, the focussing effect of Weyl curvature is that of a purely astigmatic lens, where there is positive focussing in one spatial direction perpendicular to the ray direction and an equal negative focussing in the corresponding spatial direction at right angles to this one.

These two effects are separately exhibited in the bending of light rays by the sun. Imagine the sun to be transparent and without refractive index – we are concerned only with the effect of the sun's gravitational field on idealized light rays (null geodesics) – and consider the apparent displacements, as viewed from the earth, in the star field behind and around the sun's disc. The Ricci curvature due to the positive energy density of the sun's material causes an apparent magnification of the star field behind the disc (effective positive lens) whereas the Weyl curvature in the region surrounding the sun's material causes an astigmatic distortion in the appearances of the star field surrounding the disc. If we imagine a small circular pattern in the star field behind the centre of the disc then, due to the symmetry of the light-bending effect, this pattern gets displaced symmetrically outwards to become a slightly larger circle, in accordance with the anastigmatic positive focussing of Ricci curvature. Now imagine, instead, a small circular pattern in the star field just outside the sun's disc. The image gets displaced outwards, remaining tangent to two lines drawn out from the centre of the disc. Thus the image gets stretched in the tangential direction; and it also gets squashed in the radial direction because the displacement is smaller the farther out from the centre is the point under consideration. Examining the details of this distortion a little more closely, and bearing in mind the fact that the amount of the displacement, if small, is inversely proportional to the impact parameter (i.e. to the distance of the image out from the centre of the sun's disc), we find that the *area* of the elliptical image is equal to that of the original undisplaced circle. This fact characterizes the purely astigmatic focussing due to the sun's vacuum gravitational field.

4. *The Non-locality of Gravitational Energy*

The above considerations referred only to the lowest-order effects of lensing. If that were the whole story, then there could be no enhancement of a star's image by the (vacuum) curvature field of an intervening massive body. One needs a net area *increase* to achieve magnification. This is actually is an observational (and, indeed, now observed) effect of microlensing. (See Kumar Chitre's account [7], in this volume, for a description of various ways in which gravitational lensing is having a growing importance in observational astronomy). There is indeed a net magnification as a result of the focussing due to Weyl curvature, but this is a second-order effect of non-linearities in the way that focussing power adds up. A closely related matter is the issue of the energy in a gravitational field, and the essentially non-local character of this energy. Let us examine this relationship next [6].

We have seen that the Ricci curvature due to a positive energy density of matter acts as a positive (anastigmatic) lens along a light ray. More correctly, it is the energy-tensor component in the null direction of the ray – determined, say, by the null vector l^a – which determines the degree of positive focussing :

$$T_{ab}\, l^a l^b = -\frac{1}{8\pi G} R_{ab}\, l^a l^b$$

For weak curvatures (or for regions of curvature that are all close to each other along the ray) the focussing powers of different patches of Ricci curvature essentially add, but for strong curvatures (or widely separated regions of curvature) there is a non-linear non-local contribution. This is simply just a feature of ordinary lens optics. The *power* P of a lens is the reciprocal of its focal length f :

$$\mathcal{P} = \frac{1}{f}$$

For a pair of weak (or close) lenses along the same ray, the powers simply add, but in general there is a non-linear correction term. Thus, the total power P arising from two lenses of powers P_1 and P_2 separated by a distance d is :

$$P = P_1 + P_1 - P_1P_2d$$

We can understand the term $-P_1P_2d$ as representing the (negative) potential energy in the gravitational field, in the case of two parallel flat plates of matter (Ricci curvature) separated by a distance d. In this essentially one-dimensional situation, the gravitational potential energy (per unit surface area) is proportional to the distance between the plates.

This is basically the case of two anastigmatic lenses, but we can also apply the above formula to astigmatic lenses, if we use it separately in each plane through the ray direction. Let us consider the case of a pair of parallel impulsive plane gravitational waves (delta function in the Weyl curvature, with flat space on either side), where the amplitude of the second wave is exactly opposite to that of the first. If the second wave were to follow immediately after the first one, then the two would cancel out; but suppose that they are separated by a distance d. In a principal plane through the ray direction (this being taken perpendicular to the planes of the waves) we have a focussing power p for the first wave and −p for the second – or, in the case of the opposite principal plane, −p for the first wave and p for the second. In either case we get

$$P = p^2d$$

for the total power. This shows, that in a certain sense, a pair of purely astigmatic oppositely oriented Weyl curvature lenses acts like a positively focussing Ricci lens whose power is proportional to the distance between them. This reflects the non-locality of gravitational energy (here, energy per unit surface area, as measured by focussing power), the energy not lying in (delta-function) Weyl curvature of the waves themselves but - if it lies anywhere at all - in the region of flat space between the waves.

These simple considerations can be made more precise and comprehensive, in terms of the energy-flux formula of Bondi and Sachs [8,9]. Just as above, it turns out that there is an essential non-locality and slipperiness in the energy carried by gravitational waves. This energy is nevertheless *real*, as is made manifest in the precise measurements of the binary pulsar PSR 1913+16 clearly show. Likewise, there is an inherent non-locality (and slipperiness) in the concept of gravitational *potential* energy in general relatively. These issues will have importance to us in the final section.

5. *Time-Irreversibility in Quantum Gravity*

We have seen in §2 that the space-time singularities of gravitational collapse and cosmology seem to exhibit an essential time-asymmetry, of the nature of that expressed in the Weyl curvature hypothesis. If we believe that it is "quantum gravity" - whatever that appropriate union of quantum mechanics and general relativity might turn out to be - that governs the structure of these space-time singularities (or of the neighbouring non-singular space-time), then we must believe that this "quantum gravity" will have to be a grossly time-asymmetric theory. This is hard to accept if by "quantum gravity" we mean merely a theory that results from the application of standard principles of quantum (field) theory to standard (or non-standard) general relativity. The equations of general relativity and the unitary operation **U** of standard quantum theory are both time-symmetric, and it is hard to see how a sufficiently extreme form of symmetry-breaking could give something as time-asymmetrical as the Weyl curvature hypothesis – especially since violations not only of T, PT, and CT are required, but also of CPT. This strongly suggests that our sought-for quantum gravity theory should involve also the time asymmetric operation **R** of state-vector reduction.

In fact this link can be made more explicit. Imagine a simple quantum-mechanical thought experiment, where there is a source S which emits photons from time to time, aimed at a detector D. Placed between the two is a half-silvered mirror M, tilted at an angle to the line SD, so that a photon from S, if it is reflected at M, would be absorbed at the laboratory floor at a point F. We can pose the following question :

> "When a photon is emitted at S, what is the probability that it is received at D?"

The correct answer is obtained by taking the squared modulus of the amplitude for the photon to take the path SMD. The alternative path for the photon is the reflected one SMF, whose amplitude, like that of the first path, has squared modulus $\frac{1}{2}$, giving the probability that the photon is absorbed at F. These probabilities are computed using the standard "squared-modulus rule" that characterizes the application of the procedure **R** of state-vector reduction.

Now let us consider a less usual form of question :

"When a photon is received at D, what is the probability that it was emitted at S?"

We might try to answer this question by using the same quantum-mechanical rules as before, but now in a reversed direction in time. There are two photon paths to consider, namely SMD, as before, and another path CMD which represents the photon being emitted at a certain point C on the laboratory ceiling, from which the photon could be reflected at M to reach the detector D. Examining the amplitudes for these two alternatives, we again find that their respective squared moduli are $\frac{1}{2}$, $\frac{1}{2}$. However, it is clear that these cannot now be interpreted as the respective probabilities for these two alternatives. If the detector D receives a photon, then it is almost certain that it was emitted at the source S (probability ~ 1) and that it was not ejected from the ceiling at C (probability ~ 0).

It is evident from the above that in any particular physical situation there is no need for the forward *probabilities* to blance the backward ones – even though the forward and backward *amplitudes* must balance (from unitarity). However, one can provide general phase-space arguments [10,11] to support the contention that there must be an *overall* balance between the forward and backward probabilities, when the totality of all possible physical situations is considered. The forward probabilities are governed by the squared -modulus rule demanded by **R**. However, the backward probabilities are determined by the *second law of thermodynamics* (the cold laboratory ceiling being very unlikely to emit a sufficiently energetic photon at C in order to activate D).

Now we have seen that the second law ultimately arises from the very special nature of the initial singular state of the universe, this being (presumably) a consequence of a yet-undiscovered (time-asymmetric) theory of quantum gravity. Since it is thus ultimately quantum gravity that determines the backward probabilities, it should also be this same quantum gravity that governs the forward probabilities, in order that the required overall balance can be achieved. This strongly suggests that **R** should be a quantum-gravity effect.

6. *A Gravitational Criterion for State Reduction*

Various authors [12,13,14] have suggested, from time to time that quantum state reduction should be a gravitational phenomenon. The strongest promoters of this idea have been the Hungarian school headed by Karolyhazy. Quite recently, Diosi [15] proposed a gravitational reduction scheme that followed similar lines to the non-gravitational reduction proposal of Ghirardi, Rimini, and Weber [16] and also Pearle [17]. Difficulties with Diosi's scheme were pointed out by Ghirardi, Grassi, and Rimini [18] who put forward a remedy that, however, removed the entirely gravitational nature of Diosi's proposal.

The gravitational proposal that I am making here is a little different from these, and is significantly different also from another suggestion [11] that I have been arguing for over the past several years. My present proposal falls short of being an actual theory, but it provides a usable criterion for a state-vector collapse rate in many situations.

Let us consider a situation similar to that described above, where a photon source emits a single photon whose wave function becomes split into two portions after it impinges on a half-silvered mirror. One portion encounters a detector - whose nature I shall describe in a moment - whereas the other leaves the detector undisturbed. What is the nature of this detector? Let us suppose that the measurement is registered by the displacement of a massive object - say a "lump" - from one place to another. (If one is inclined to believe that the actual "measurement" is likely to take place earlier in the detection process, this detector could be taken to consist, ideally, of just a lump so delicately poised at the edge of a precipice that if it is struck by a photon, then it would fall over the edge to the bottom). Since the photon wave function becomes a superposition of two beams, standard quantum mechanics would maintain that the position of the lump should itself become a superposition. My present proposal [19,20,21], however, is that this superposition is to be regarded as unstable - analogous to an unstable nucleus - having two decay modes, one being to the state in which the lump is in its original position (in conjunction with the reflected photon state) and the other in which the lump is in the displaced position. The characteristic decay time for each of these two modes (where I am assuming - as in this example - that the amplitudes for the two superposed alternatives are roughly equal in magnitude) is to be of the order of $\hbar E^{-1}$, where E is the gravitational self-energy of the difference between the two mass distributions under consideration. Thus, in this case, we take the mass distribution consisting of the lump in its initial location and substract from it the mass distribution of the mass in its displaced location. The quantity E is the gravitational self-energy of this arrangement of positive and negative masses.

We can think of E in a slightly different way in cases, such as the above, where the two mass distributions are simply rigid displacements of one another. Here E is simply the energy that it would take to pull the lump away to its displaced location, in the gravitational field provided by the lump in its original location. For a spherical lump of uniform density, the energy required to move the lump from the original location to a place where it just ceases to overlap its original location is of the same general order as that which would be required to displace it from there all the way out to infinity, this being of the order of $E = Gm^2/a$. In such situations the reduction time is, therefore,

$$\hbar a/Gm^2$$

approximately,where a is the radius and m the mass of the lump.

We can see how this applies in various examples. For a lump consisting of a single proton (taking its "radius" to be its de Broglie radius) we get a reduction time of about 10^6 years. For a speck of water one micron in radius, we get a reduction time of about a tenth of a second. Had the droplet's radius been 10 microns or so, we would get a reduction time of about 10^{-6} seconds. Had its radius been 1/10 micron, then reduction would take a few hours.

In these considerations I have ignored the granular nature of the lump's material and, more importantly, the fact that the environment would itself become part of the superposed state. The granular nature of the lump itself turns out not to be important; but the granular nature of the environment *is* important, and as soon as this environment becomes significantly entangled with the lump's state, reduction would be likely to occur first because of mass displacements in that environment rather than in the mass itself, leading to a much shorter reduction time than that computed explicitly above.

7. *Tie-up with Non-locality of Gravitational Energy*

The above general proposal gains some support from the following consideration. Standard application of Heisenberg's uncertainty principle to unstable particles tells us that there is an uncertainty in the initial mass-energy which is of the order of $\hbar$ times the reciprocal of the lifetime. In the context of the proposal set forth in the previous section, therefore, there should be a mass-energy uncertainty in the state of superposed mass distributions which is of the order of the gravitational self-energy in the difference between these mass distributions.

Now, in Newtonian theory this gravitational self-energy can be defined unambiguously, but we must bear in mind that it is general relativity to which we should turn in connection with fundamental issues such as those under consideration here. We are examining a state which involves the superposition of two slightly different space-time geometries, and there is no meaning to the concept of gravitational field energy in such a superposition, or in a "difference" between the mass distributions - even non-locally - unless one can assign a precise mapping between the points of two space-times to be superposed. Moreover, the essential *non-locality* of gravitational field energy, as encountered in §4 is an unavoidable consequence of the principles of general relativity.

As yet, these considerations are heuristic rather than clear cut, but they strongly indicate to me that within a more comprehensive and precisely formulated theory of gravitationally induced quantum state reduction, there will be a dovetailing of the field energy problems of general relativity with the energy problems that have been encountered in quantum state reduction [18], and it is to be hoped that a fully consistent scheme along these lines can thereby be achieved.

REFERENCES

1. E.P.Wigner (1960) The unreasonable effectiveness of Mathematics in the physical sciences, Commun.Pure Appl.Math. 13, 1-14.
2. J. H.Taylor (1993) Testing relativistic gravity with binary and millisecond pulsars, in *General Relativity and Gravitation 1992*, edited by R.J.Gleiser, C.N.Kozameh, and O.M.Moreschi (I.O.P., Bristol).
3. S.W.Hawking & R.Penrose (1970) The singularities of gravitational collapse and cosmology, Proc.Roy.Soc.(Lond.) A314, 529-48.
4. R.Penrose (1979) Singularities and time-asymmetry,in General Relativity, an Einstein Centenary Survey, edited by S.W.Hawking and W.Israel (Camb. U.Press, Cambridge).
5. R. Penrose and W. Rindler (1984), *Spinors and Space-time Vol.I* (Camb U. Press ,Cambridge).
6. R.Penrose (1966) General-relativistic energy flux and elementary optics, in*Perspectives in Geometry and Relativity*, edited by B.Hoffman (Indiana U. Press, Bloomington).
7. K.Chitre (1994) This volume.
8. H.Bondi, M.C.J.van der Burg and A.W.K. Metzner (1962) *Gravitational waves in general relativity* VII ; Waves from axi-symmetric isolated systems, Proc. Roy.Soc.(Lond.) A269, 21-52.
9. R.K.Sachs (1962) *Gravitational waves in general relativity* VIII; Waves in asymptotically flat space-time, Proc.Roy Soc.(Lond.)A270, 103-26.

10. R.Penrose (1981) Time-asymmetry and quantum gravity , in *Quantum Gravity 2, a Second Oxford Symposium* , edited by C.J.Isham, R.Penrose and D.W.Sciama (Oxford U.Press, Oxford).

11. R.Penrose (1989)*The Emperor's New Mind*, (Oxford.U.Press,Oxford).

12. F.Karolyhazy (1966) Gravitation and quantum mechanics of macroscopic bodies, Nuovo Cim.A42, 390-402.

13. F.Karolyhazy, A.Frenkel and B. Lukacs (1986) On the possible role of gravity on the reduction of the wavefunction, in *Quantum Concepts in Space and Time*, edited by R.Penrose and C.J.Isham (Oxford U. Press, Oxford).

14. A. B.Komar (1969) Qualitative features of quantized gravitation, Int. J. Theor.Phys.2, 157-60.

15. L.Diosi (1989) Models for universal reduction of macroscopic quantum fluctuations, Phys.Rev.A40, 1165-74.

16. G.C.Ghirardi, A.Rimini and T.Weber (1986) Unified dynamics for microscopic and macroscopic systems, Phys.Rev. D34, 470.

17. P. Pearle (1989) Combining stochastic dynamical statevector reduction with spontaneous localization, Phys.Rev.A39, 2277-89.

18. G.C.Ghirardi, R.Grassi and A.Rimini (1990) Continuous-spontaneous-reduction model involving gravity, Phys.Rev.A42, 1057-64.

19. R. Penrose (1993) Gravity and Quantum Mechanics in *General Relativity and Gravitation* 1992, edited by R.J.Gleiser, C.N.Kozameh and O.M. Moreschi (I.O.P., Bristol).

20. R. Penrose (1994) Non-locality and objectivity in quantum state reduction , in *Fundamental Aspects of Quantum Theory*, editrd by J. Anandan and J.L. Safko (World Scientific, Singapore).

21. R. Penrose (1994) *Shadows of the Mind* (Oxford U. Press, Oxford).

243

CONSCIOUSNESS INVOLVES NONCOMPUTABLE INGREDIENTS

The third culture: Beyond the scientific revolution. Ed. J. Brockman, 239–47. Simon and Schuster, New York, 1995

594 Consciousness involves noncomputable ingredients. The third culture: *Beyond the scientific revolution*. Eds J. Brockman, 239–47. 1995

Roger Penrose: My main technical interest is in twistor theory—a radical approach to space and time—and, in particular, how to fit it in with Einstein's general relativity. There's a major problem there, in which some progress was made a few years ago, and I feel fairly excited about it. It's ultimately aimed at finding the appropriate union between general relativity and quantum theory.

When I was first seriously thinking of getting into physics, I was thinking more in terms of quantum theory and quantum electrodynamics than of relativity. I never got very far with quantum theory at that stage, but that was what I started off trying to do in physics. My Ph.D. work had been in pure mathematics. I suppose my most quoted paper from that period was on generalized inverses of matrices, which is a mathematical thing that physicists hardly ever mention. Then there were the nonperiodic tilings, which relate to quasi crystals, and therefore to solid-state physics to some degree. Then there's general relativity. What I suppose I'm best known for in that area are the singularity theorems that I worked on along with Stephen Hawking. I knew him when he was Dennis Sciama's graduate student; I've known him for a long time now. But the main things I've done in relativity apart from that have to do with spinors and with asymptotic structure of spacetimes, relating to gravitational radiation.

I believe that general relativity will modify the structure of quantum mechanics. Whereas people usually think that in order to unite quantum theory with gravity theory you should apply quantum mechanics, unmodified, to general relativity, I believe that the rules of quantum theory must themselves be modified in order for this union to be successful.

There's a connection between this area of physics and consciousness, in my opinion, but it's a bit roundabout; the arguments are negative. I argue that we shall need to find some noncomputational physical process if we're ever to explain the effects of consciousness. But I don't see it in any existing theory. It seems to me that the only place where noncomputability can possibly enter is in what is called "quantum measurement." But we need a new theory of quantum measurement. It must be a noncomputable new theory. There is scope for this, if the new theory involves changes in the very structure of quantum theory, of the kind that could arise when it's appropriately united with general relativity. But this is something for the distant future.

Why do I believe that consciousness involves noncomputable ingredients? The reason is Gödel's theorem. I sat in on a course when I was a research student at Cambridge, given by a logician who made the point about Gödel's theorem that the very way in which you show the formal unprovability of a certain proposition also exhibits the fact that it's true. I'd vaguely heard about Gödel's theorem—that you can produce statements that you can't prove using any system of rules you've laid down ahead of time. But what was now being made clear to me was that as long as you believe in the rules you're using in the first place, then you must also believe in the truth of this proposition whose truth lies beyond those rules. This makes it clear that mathematical understanding is something you can't formulate in terms of rules. That's the view which, much later, I strongly put forward in my book *The Emperor's New Mind*.

There are possible loopholes to this use of Gödel's theorem, which people can pick on, and they often do. Most of these counterarguments are misunderstandings. Dan Dennett makes genuine points, though, and these need a little more work to see why they still don't get around the Gödel argument. Dennett's case rests on the contention that we use what are called "bottom-up" rather than "top-down" algorithms in our thinking—here, mathematical thinking.

A top-down algorithm is specific to the solution of some particular problem, and it provides a definite procedure that is known to solve that problem. A bottom-up algorithm is one that is not specific to any particular problem but is more loosely organized, so that it learns by experience and gradually improves, eventually giving a good solution to the problem at hand. Many people have the idea that bottom-up systems rather than top-down, programmed algorithmic systems are the way the brain works. I apply the Gödel argument to bottom-up systems too, in my most recent book, *Shadows of the Mind*. I make a strong case that bottom-up systems also won't get around the Gödel argument. Thus, I'm claiming, there's something in our conscious understanding that simply isn't computational; it's something different.

A lot of what the brain does you could do on a computer. I'm not saying that all the brain's action is completely different from what you do on a computer. I am claiming that the actions of consciousness are something different. I'm not saying that consciousness is beyond physics, either—although I'm saying that it's beyond the physics we know now.

The argument in my latest book is basically in two parts. The first part shows that conscious thinking, or conscious understanding, is something different from computation. I'm being as rigorous as I can about that. The second part is more exploratory and tries to find out what on earth is going on. That has two ingredients to it, basically.

My claim is that there has to be something in physics that we don't yet understand, which is very important, and which is of a noncomputational character. It's not specific to our brains; it's out there, in the physical world. But it usually plays a totally insignifi-cant role. It would have to be in the bridge between quantum and classical levels of behavior—that is, where quantum measurement comes in.

Modern physical theory is a bit strange, because one has two levels of activity. One is the quantum level, which refers to small-scale phenomena; small energy differences are what's relevant. The other level is the classical level, where you have large-scale phenomena, where the roles of classical physics—Newton, Maxwell, Einstein—operate. People tend to think that because quantum mechanics is a more modern theory than classical physics, it must be more accurate, and therefore it must explain classical physics if only you could see how. That doesn't seem to be true. You have two scales of phenomena, and you can't deduce the classical behavior from the quantum behavior any more than the other way around.

We don't have a final quantum theory. We're a long way from that. What we have is a stopgap theory. And it's incomplete in ways that affect large-scale phenomena, not just things on the tiny scale of particles.

Current physics ideas will survive as limiting behavior, in the same sense that Newtonian mechanics survives relativity. Relativity modifies Newtonian mechanics, but it doesn't really supplant it. Newtonian mechanics is still there as a limit. In the same sense, quantum theory, as we now use it, and classical physics, which includes Einstein's general theory, are limits of some theory we don't yet have. My claim is that the theory we don't yet have will contain noncomputational ingredients. It must play its role when you magnify something from a quantum level to a classical level, which is what's involved in "measurement."

The way you treat this nowadays, in standard quantum theory, is to introduce randomness. Since randomness comes in, quantum theory is called a probabilistic theory. But randomness only comes in when you go from the quantum to the classical level. If you stay down at the quantum level, there's no randomness. It's only when you magnify something up, and you do what people call "make a measurement." This consists of taking a small-scale quantum effect and magnifying it out to a level where you can see it. It's only in that process of magnification that probabilities come in. What I'm claiming is that whatever it is that's really happening in that process of magnification is different from our present understanding of physics, and it is not just random. It is noncomputational; it's something essentially different.

This idea grew from the time when I was a graduate student, and I felt that there must be something noncomputational going on in our thought processes. I've always had a scientific attitude, so I believed that you have to understand our thinking processes in terms of science in some way. It doesn't have to be a science that we understand now. There doesn't seem to be any place for conscious phenomena in the science that we understand today. On the other hand, people nowadays often seem to believe that if you can't put something on a computer, it's not science.

I suppose this is because so much of science is done that way these days; you simulate physical activity computationally. People don't realize that something can be noncomputational and yet perfectly scientific, perfectly mathematically describable. The fact that I'm coming into all this from a mathematical background makes it easier for me to appreciate that there are things that aren't computational but are perfectly good mathematics.

When I say "noncomputational" I don't mean random. Nor do I mean incomprehensible. There are very clear-cut things that are noncomputational and are known in mathematics. The most famous example is Hilbert's tenth problem, which has to do with solving algebraic equations in integers. You're given a family of algebraic equations and you're asked, "Can you solve them in whole numbers? That is, do the equations have integer solutions?" That question—yes or no, for any particular example—is not one a computer could answer in any finite amount of time. There's a famous theorem, due to Yuri Matiyasevich, which proves that there's no computational way of answering this question in general. In particular cases, you might be able to give an answer by means of some algorithmic procedure. However, given any such algorithmic procedure, which you know doesn't give you wrong answers, you can always come up with an algebraic equation that will defeat that procedure but where you know that the equation actually has no integer solutions.

Whatever understandings are available to human beings, there are—in relation particularly to Hilbert's tenth problem—things that can't be encapsulated in computational form. You could imagine a toy universe that evolved in some way according to Hilbert's tenth problem. This evolution could be completely deterministic yet not computable. In this toy model, the future would be mathematically fixed; however, a computer could not tell you what this future is. I'm not saying that this is the way the laws of physics work at some level. But the example shows you that there's an issue. I'm sure the real universe is much more subtle than that.

The Emperor's New Mind served more than one purpose. Partly I was trying to get a scientific idea across, which was that noncomputability is a feature of our conscious thinking, and that this is a perfectly reasonable scientific point of view. But the other part of it was educational, in a sense. I was trying to explain what modern physics and modern mathematics is like.

Thus, I had two quite different motivations in writing the book. One was to put a philosophical point of view across, and the other was that I felt I wanted to explain scientific things. For quite a long time, I'd felt that I did want to write a book at a semipopular level to explain certain ideas that excited me—ideas that weren't particularly unconventional—about what science is like. I had it in the back of my mind that someday I would do such a thing.

It wasn't until I saw a BBC "Horizon" program, in which Marvin Minsky and various people were making some rather extreme and outrageous statements, that I was finally moved to write the book. I felt that there was a point of view which was essentially the one I believe in, but which I had never seen expressed anywhere and which needed to be put forward. I knew that this was what I should do. I would write this book explaining a lot of things in science, but this viewpoint would give it a focus. Also it had to be a book, because it's cross-disciplinary and not something you could express very well in any particular journal.

I suppose what I was doing in that book was philosophy, but somebody complained that I hardly referred to a single philosopher—which I think is true. That's because the questions that interest philosophers tend to be rather different from those that interest scientists; philosophers tend to get involved in their own internal arguments.

When I argue that the action of the conscious brain is noncomputational, I'm not talking about quantum computers. Quantum computers are perfectly well-defined concepts, which don't involve any change in physics; they don't even perform noncomputational actions. Just by themselves, they don't explain what's going on in the conscious actions of the brain. Dan Dennett thinks of a quantum computer as a skyhook, his term for a miracle. However, it's a perfectly sensible thing. Nevertheless, I don't think it can explain the way the brain works. That's another misunderstanding of my views. But there could be some element of quantum computation in brain action. Perhaps I could say something about that.

One of the essential features of the quantum level of activity is that you have to consider the coexistence of various different alternative events. This is fundamental to quantum mechanics. If X can happen, and if Y can happen, then any combination of X and Y, weighted with complex coefficients, can also occur. According to quantum mechanics, a particle can have states in which it occupies several positions at once. When you treat a system according to quantum mechanics, you have to allow for these so-called superpositions of alternatives.

The idea of a quantum computer, as it's been put forward by David Deutsch, Richard Feynman, and various other people, is that the computations are the things that are superposed. Rather than your computer doing one computation, it does a lot of them all at once. This may be, under certain circumstances, very efficient. The problem comes at the end, when you have to get one piece of information out of the superposition of all those different computations. It's extremely difficult to have a system that does this usefully.

It's pretty radical to say that the brain works this way. My present view is that the brain isn't exactly a quantum computer. Quantum actions are important in the way the brain works, but the brain's noncomputational actions occur at the bridge from the quantum to the classical level, and that bridge is beyond our present understanding of quantum mechanics.

The most promising place by far to look for this quantum-classical borderline action is in recent work on microtubules by Stuart Hameroff and his colleagues at the University of Arizona. Eukaryotic cells have something called a cytoskeleton, and parts of the cytoskeleton consist of these microtubules. In particular, microtubules inhabit neurons in the brain. They also control one-celled animals, such as parameciums and amoebas, which don't have any neurons. These animals can swim around and do very complicated things. They apparently learn by experience, but they're not controlled by nervous systems; they're controlled by another kind of structure, which is probably the cytoskeleton and its system of microtubules.

Microtubules are long little tubes, a few nanometers in diameter. In the case of the microtubules lying within neurons, they very likely extend a good deal of the length of the axons and the dendrites. You find them from one end of the axons and dendrites to the other. They seem to be responsible for controlling the strengths of the connections between different neurons. Although at any one moment the activity of neurons could resemble that of a computer, this computer would be subject to continual change in the way it's "wired up," under the control of a deeper level of structure. This deeper level is very probably the system of microtubules within neurons.

Their action has a lot to do with the transport of neurotransmitter chemicals along axons, and the growth of dendrites. The neurotransmitter molecules are transported along the microtubules, and these molecules are critical for the behavior of the synapses. The strength of the synapse can be changed by the action of the microtubules. What interests me about the microtubules is that they're tubes, and according to Hameroff and his colleagues there's a computational action going along on the tubes themselves, on the outside.

A protein substance called tubulin forms interpenetrating spiral arrangements constituting the tubes. Each tubulin molecule can have two states of electric polarization. As with an electronic computer, we can label these states with a 1 and a 0. These produce various patterns along the microtubules, and they can go along the tubes in some form of computational action. I find this idea very intriguing.

By itself, a microtubule would just be a computer, but at a deeper level than neurons. You still have computational action, but it's far beyond what people are considering now. There are enormously more of these tubulins than there are neurons. What also interests me is that within the microtubules you have a plausible place for a quantum-oscillation activity that's isolated from the outside. The problem with trying to use quantum mechanics in the action of the brain is that if it were a matter of quantum nerve signals, these nerve signals would disturb the rest of the material in the brain, to the extent that the quantum coherence would get lost very quickly. You couldn't even *attempt* to build a quantum computer out of ordinary nerve signals, because they're just too big and in an environment that's too disorganized. Ordinary nerve signals have to be treated classically. But if you go down to the level of the microtubules, then there's an extremely good chance that you can get quantum-level activity inside them.

For my picture, I need this quantum-level activity in the microtubules; the activity has to be a large-scale thing that goes not just from one microtubule to the next but from one nerve cell to the next, across large areas of the brain. We need some kind of coherent activity of a quantum nature which is weakly coupled to the computational activity that Hameroff argues is taking place along the microtubules.

There are various avenues of attack. One is directly on the physics, on quantum theory, and there are certain experiments that people are beginning to perform, and various schemes for a modification of quantum mechanics. I don't think the experiments are sensitive enough yet to test many of these specific ideas. One could imagine experiments that might test these things, but they'd be very hard to perform.

On the biological side, one would have to think of good experiments to perform on microtubules, to see whether there's any chance that they do support any of these large-scale quantum coherent effects. When I say "quantum coherent effects," I mean things a bit like superconductivity or superfluidity, where you have quantum systems on a large scale.

244

BEYOND THE DOUBTING OF A SHADOW: A REPLY TO COMMENTARIES ON *SHADOWS OF THE MIND*

Psyche **2** (23) 1996

As with Chapters 196 and 225, this is a response to critical commentaries, but this time concerning my book *Shadows of the Mind* (Oxford University Press, 1994). I had to work hard to try to address the main substantive points raised. However, there is an additional issue whose significance I have come to appreciate only rather recently. Chalmers's 'succinct summary' of my 'central new argument', given on page five, beginning of Section 3., under (A) actually differs in a significant way from my own argument, with regard to the computational implementation of the assertion 'I am F' or 'I know that I am F', referred to in 3.3 and 3.4. This matter is addressed in my article 'Gödel, the mind, and the laws of physics' which is due to appear in the proceedings of the Vienna Conference, *Horizons of Truth*, Gödel Centenary 2006, An International Symposium Celebrating the 100th Birthday of Kurt Gödel, 27–29 April 2006, *Festsaal* of the University of Vienna.

Beyond the Doubting of a Shadow

A Reply to Commentaries on *Shadows of the Mind*

Roger Penrose
Mathematical Institute
24-29 St. Giles
Oxford OX1 3LB
U.K.

Copyright (c) Roger Penrose 1996

PSYCHE, 2(23), January 1996
http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html

KEYWORDS: artificial intelligence, free will, Gödel's theorem, mathematics, microtubules, Platonism, quantum mechanics.

REPLIES TO:

1. Bernard J. Baars: Can physics provide a theory of consciousness?
2. David J. Chalmers: Minds, machines, and mathematics
3. Solomon Feferman: Penrose's Gödelian argument
4. Stanley A. Klein: Is quantum mechanics relevant to understanding consciousness?
5. Tim Maudlin: Between the motion and the act....
6. John McCarthy: Awareness and understanding in computer programs
7. Daryl McCullough: Can humans escape Gödel?
8. Drew McDermott: [STAR] Penrose is wrong
9. Hans Moravec: Roger Penrose's gravitonic brains

CONTENTS

1. General remarks
2. Some technical slips in *Shadows*
3. The central new argument of *Shadows*
4. The "bare" Gödelian case
5. Gödel's "theorem-proving machine"
6. The issue of errors
7. The "unknowability" issue
8. AI and MJC
9. Mathematical Platonism
10. What has Gödel's theorem to do with physics?
11. How could physics actually help?
12. State-vector reduction
13. Free will
14. Some remarks on biology
15. What is consciousness?

1. General Remarks

1.1 I am glad to have this opportunity to address some of the criticisms that have been aimed at arguments in my book *Shadows of the Mind* (henceforth *Shadows*). I hope that in the following remarks I am able to remove some of the confusions and misunderstandings that still surround the arguments that I tried to make in that book - and also that we may be able to move forward from there.

1.2 In the accompanying PSYCHE articles, the great majority of the commentators' specific criticisms have been concerned with the purely logical arguments given in Part 1 of *Shadows*, with comparatively little reference being made to the physical arguments given in Part 2 - and virtually none at all to the biological ones.<1> This is not unreasonable if it is regarded that the entire rationale for my physical and biological arguments stands or falls with my purely logical arguments. Although I do not entirely agree with this position - since I believe that there are strong motivations from other directions for the kinds of physical and biological action that I have been promoting in *Shadows* - I am prepared to go along with it for the moment. Thus, most of my remarks here will be concerned with the implications of Gödel's theorem, and with the claims made by many of my critics that my arguments do not actually establish that there must be a noncomputational ingredient in human conscious thinking.

1.3 In replying to these arguments, I should first point out that, very surprisingly, almost none of the commentators actually addresses what I had regarded as the central (new) core argument against the computational modelling of mathematical understanding! Only Chalmers actually draws attention to it, and comments in detail on this argument, remarking that "most commentators seem to have missed it".<2> Chalmers also remarks that "it is unfortunate that this argument was so deeply buried". I apologize if this appears to have been the case; but I am also very puzzled, since its essentials are summarized in the final arguments of "Reductio ad absurdum - a fantasy dialogue", which is the section of *Shadows* (namely Section 3.23) that readers are particularly directed towards. This section is referred to also by McDermott and by Moravec, but neither of these commentators actually addresses this central argument explicitly, and nor do any of the other commentators. This is particularly surprising in the case of McCullough, as he is concerned with some of the subtleties of the logic involved, and also of Feferman, in view of his very carefully considered logical discussion.

1.4 It would appear, therefore, that I have an easy solution to the problem of replying to all nine commentators. All I need do is show why the ingenious argument put forward by Chalmers (based partly on McCullough's very general considerations) as a counter to my central argument is in fact (subtly) invalid! However, I am sure that this mode of procedure would satisfy none of the other commentators, and many of them also have interesting other points to make which need commenting upon. Accordingly, in the following remarks, I shall attempt to address all the serious points that they do bring up. My reply to this main argument of Chalmers (partly dependent upon that of McCullough) will be given in Section 3, but it will be helpful first to precede this by addressing, in Section 2, the significant logical points that are raised by Feferman in his careful commentary.

2. Some Technical Slips in *Shadows*

2.1 Feferman quite correctly draws attention to some inaccuracies in *Shadows* with regard to certain logical technicalities. The most significant of these (in fact, the only really significant one for my actual arguments) concerns a misunderstanding on my part with regard to the assertion of omega-consistency of a formal system F, which I had chosen to denote by the symbols Omega(F), and its relation to Gödel's first incompleteness theorem. (As it happens, two others before Feferman had also pointed out this particular error to me.) As Feferman says, the assertion that some particular formal system is "omega-consistent" is certainly not of the form of a PI_1-sentence (i.e. not of the form of an assertion: "such-and-such a Turing computation never halts" - I call these "P-sentences" from here on). This much I should have been (and essentially was) aware of, despite the fact that in the first two printings of *Shadows*, p.96 I made the assertion that Omega(F) is a P-sentence. The fact of the matter was that I had somehow (erroneously) picked up the belief that the statement that Gödel originally exhibited in his famous first incompleteness theorem was *equivalent* to the omega-consistency of the formal system in question, not that it merely *followed* from this omega-consistency. Accordingly, I had imagined that for some technical reason I did not know of, this omega-consistency must actually be equivalent (for sufficiently extensive systems F) to the particular assertion "C_k(k)" that I had exhibited in Section 2.5, when the rules of the formal system F are translated into the algorithm A. Accordingly, I had mistakenly believed that Omega(F) must, for some subtle reason (unknown to me), be equivalent to the P-sentence C_k(k) (at least for sufficiently extensive systems F).

2.2 This error affects none of the essential arguments of the book but it is unfortunate that in various parts of Chapter 3, and most particularly in the "fantasy dialogue" in Section 3.23, the notation "Omega(F)" is used in circumstances where I had intended this to stand for the actual P-sentence C_k(k). In later printings of *Shadows*, this error has been corrected: I use the Gödel sentence G(F) (which asserts the consistency of F and is a P-sentence) in place of Omega(F). It is in any case much more appropriate to use G(F) in the arguments of Chapter 3, rather than Omega(F), and I agree with Feferman that the introduction of "Omega(F)" was essentially a red herring. In fact, the presentation in *Shadows* would have usefully simplified if omega-consistency had not even been mentioned.

2.3 The next most significant point of inaccuracy - or rather imprecision - in *Shadows* that Feferman brings up is that there is a discrepancy between different notions of the term "sound" that I allude to in different parts of the book. (This is actually quite an important issue, in relation to some of the discussion to follow, and I shall need to return to it later in Section 3.) His point is, essentially, that in some places I need make use of the soundness of a formal system only in the limited sense of its capacity to assert the truth of certain P-sentences, whereas in other places I am actually referring to soundness in a more comprehensive sense, where it applies to other types of assertion as well. I agree that I should have been more careful about such distinctions. In fact, it is the weaker notion of soundness that would be sufficient for all the "Gödelian" arguments that I actually use in Part 1 of *Shadows*, though for some of the more philosophical discussions, I had in mind soundness in a stronger sense. (This stronger sense is not needed on pp. 90-92 if omega-consistency is dropped; nor is it needed on p.112, the weaker notion of soundness now being equivalent to consistency.)

2.4 Basically, I am happy to agree with all the technical criticisms and corrections that Feferman refers to in his section discussing my treatment of the logical facts". (I should attempt a point of clarification concerning his puzzlement as to why I should make the "strange" and "trivial" assertions he refers to on p.112. No doubt I expressed myself badly. The point that I was attempting to make concerned the issue of the relationship between the formal string of symbols that constitute "G(F)" and "Omega(F)" and the actual *meanings* that these strings are supposed to represent. I was merely trying to argue that meanings are essential - a point with which Feferman strongly concurs, in his commentary.) It should be made clear that none of these corrections affects the arguments of Chapter 3 in any way (so long as Omega(F) is replaced by G(F) throughout), as Feferman himself appears to affirm in his last paragraph of the aforementioned section.

2.5 I find it unfortunate, however, that he does not offer any critique of the arguments of Chapter 3. I would have found it very valuable to have had the comments of a first-rate logician such as himself on some of the specifics of the discussions in Chapter 3. Feferman seems to be led to having some unease about the arguments presented there, not because of specific errors that he has detected, but merely because my "slapdash scholarship" may be "stretched perilously thin in areas different from [my] own expertise". A related point is made by McCarthy, McDermott and Baars in connection with my evidently inadequate referencing of the literature on AI, and on other theories that relate to consciousness, either in its computational, biological, or psychological respects.

2.6 I think that a few words of explanation, from my own vantage point, are necessary here. An ability to search thoroughly through the literature has never been one of my strong points, even in my own subject (whatever that might be!). My method of working has tended to be that I would gather some key points from the work of others and then spend most of my time working entirely on my own. Only at a much later stage would I return to the literature to see how my evolved views might relate to those of others, and in what respects I had been anticipated or perhaps contradicted. Inevitably I shall miss things and get some things wrong. The most likely source of error tends to be with second-hand information, where I might misunderstand what someone else tells me when reporting on the work of a third person. Gradually these things sort themselves out, but it takes time.

2.7 My reason for mentioning this is to emphasize that errors of the nature of those pointed out by Feferman are concerned essentially with this link of communication with the outside (scientific, philosophical, mathematical, etc.) world, and not with the internal reasonings that constitute the essential Gödelian arguments of *Shadows*. Most specifically, the main parts of Chapter 3 (particularly 3.2, 3.3 and 3.5-3.24) are entirely arguments that I thought through on my own, and are therefore independent of however "slapdash" my scholarship might happen to be! I trust that these arguments will be judged entirely on their intrinsic merits.

3. The Central New Argument of *Shadows*

3.1 Chalmers provides a succinct summary of the central new argument that I presented in *Shadows* (Section 3.16, and also 3.23 and 3.24 - but recall that my Omega(F) should be replaced by G(F) throughout Section 3.16 and 3.23). Let me repeat the essentials of Chalmers's presentation here - but with one important distinction, the significance of which I shall explain in a moment.

3.2 We try to suppose that the totality of methods of (unassailable) mathematical reasoning that are in principle humanly accessible can be encapsulated in some (not necessarily computational) sound formal system F. A human mathematician, if presented with F, could argue as follows (bearing in mind that the phrase "I am F" is merely a shorthand for "F encapsulates all the humanly accessible methods of mathematical proof"):

(A) "Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F". I perceive that it follows from the assumption that I am F that the Gödel statement G(F') would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "if I happened to be F, then G(F') would have to be true", and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that, I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F."

3.3 (Of course, one might worry about how an assertion like "I am F" might be made use of in a logical formal system. In effect, this is discussed with some care in *Shadows*, Sections 3.16 and 3.24, in relation to the Sections leading up to 3.16, although the mode of presentation there is somewhat different from that given above, and less succinct.)

3.4 The essential distinction between the above presentation and that of Chalmers is that he makes use (in the first (2) of his Section 2) of the stronger conditional assumption "I know that I am F", rather than merely "I am F", the latter being all that I need for the above. Thus, if we accept the validity of the above argument, the conclusion is considerably stronger than the "strong" conclusion that Chalmers draws ("threatening to the prospects of AI") to the effect that it "would rule out even the possibility that we could empirically discover that we were identical to some system F".

3.5 In fact, it *was* this stronger version (A) that I presented in *Shadows*, from which we would conclude that we cannot be identical to any knowable (Gödelizable) system F whatever, whether we might empirically come to believe in it or not! I am sure that this stronger conclusion would provide an even greater motivation for people (whether AI supporters or not) to find a flaw in the argument. So let me address the particular objection that Chalmers (and, in effect, also McCullough) raises against it.

3.6 The problem, according to Chalmers, is that it is contradictory to "know that we are sound". Accordingly, he argues, it would be invalid to deduce the soundness of F, let alone that of F', from the assumption "I am F". On the face of it, to a mathematician, this seems an unlikely let-out, since in all the above discussions we are referring simply to the notion of *mathematical proof.* Moreover, the "I" in the above discussion refers to an idealized human mathematician. (The problems that this notion raises, such as those referred to by McDermott, are not my concern at the moment. I shall return to such matters later; cf. Section 6.) Suppose that F indeed represents the totality of the procedures of mathematical proof that are in principle humanly accessible. Suppose, also, that we happen to come across F and actually entertain this possibility that we might "be" F, in this sense (without actually knowing, for sure, whether or not we are indeed F). Then, under the assumption that it *is* F that encapsulates all the procedures of valid mathematical proof, we must surely conclude that F is sound. The whole point of the procedures of mathematical proof is that they instil *belief.* And the whole point of the Gödel argument, as I have been employing it, is that a belief in the conclusions that can be obtained using some system H entails, also, a belief in the soundness and consistency of that system, together with a belief (for a Gödelizable H) that this consistency cannot be derived using H alone.

3.7 This notwithstanding, Chalmers and McCullough argue for an inconsistency of the very notion of a "belief system" (which, as I have pointed out above, simply means a system of procedures for mathematical proof) which can believe in itself (which means that mathematicians actually trust their proof procedures). In fact, this conclusion of inconsistency is far too drastic, as I shall show in a moment. The key issue is not that belief systems are inconsistent, or incapable of trusting themselves, but that they must be restricted as to what kind of assertion they are competent to address.

3.8 To show that "a belief system which believes in itself" need not be inconsistent, consider the following. We shall be concerned just with P-sentences (which, we recall, are assertions that specified Turing machine actions do not halt). The belief system B, in question, is simply the one which "believes" (and is prepared to assert as "unassailably perceived") a P-sentence S if and only if S happens to be true. B is not allowed to "output" anything other than a decision as to whether or not a suggested P-sentence is true or false - or else it may prattle on, as is its whim, generating P-sentences together with their correct truth values. However, as part of its internal musings, it is allowed to contemplate other kinds of thing, such as the fact that it *does*, indeed, produce only truths in its decisions about P-sentences. Of course, B is not a computational system - it is a Turing *oracle* system, as far as its output is concerned - but that should not matter to the argument. Is there anything wrong in B "believing in the soundness of B"? Nothing whatever, if we interpret this in the right way. The important thing is that B is allowed only to make assertions about P-sentences. It can use whatever procedures it likes in its internal musings, but all its *outputs* must be assertions as to the validity of particular P-sentences. If we apply the diagonal procedure that Chalmers and McCullough refer to, then we get something which is not a P-sentence, and is accordingly not allowed to be part of this belief system's output.

3.9 It may be felt that this is a pretty limited kind of "belief system", where it can make assertions only about the truth or falsity of P-sentences. Perhaps it is limited; but it is precisely a belief system of this very kind that comes into the arguments of Chapter 3 of *Shadows*. In that discussion, I was careful, in the key Section 3.16 of

Shadows, to limit the mathematical assertions under consideration to P-sentences. This avoids many difficult issues that can arise without such restrictions. However, the robots described in that section are allowed to think in very general terms - as human mathematicians may do - about non-computable systems and uncountable cardinals, etc. Nevertheless, the *-assertions under consideration must always be P-sentences, and it is *only* in relation to such sentences (as outputs) that the formal systems Q(M) and Q_M(M) are constructed. In this circumstance the argument serves to show that the robots' "belief system" cannot, after all, be a computational one, provided that it is broad enough to encompass Gödelian reasoning - which is a contradiction with the notion of "robot" that was being used.

3.10 This is not to say that the diagonalization procedure that McCullough and Chalmers refer to need apply only to computational belief systems F. As they both argue (particularly McCullough), there is no requirement that F be computational in their discussions. Indeed, in Section 7.9 of *Shadows* (which is in Part 2, so it is easy to miss, if one is concerned only with the logical arguments of that book - and neither McCullough nor Chalmers actually mention it), I explicitly referred to the fact that the Gödel-type diagonalization arguments of Part 1 will apply much more generally than merely to computational systems. For example, if Turing's oracle-computation notions are adopted, then the diagonalization procedures are quite straight-forward. However, in any specific application, it is necessary to restrict the class of sentences to which the notion of "unassailable belief" can be applied. If we do not do this, we can land in paradox, which is exactlythe situation that McCullough and Chalmers find themselves in.

3.11 Indeed, McCullough actually carries through such paradoxical reasoning in his Section 2.1, seeming to be presenting this parody of my own reasoning as though it were actually my own reasoning. This is beneath his usual standards. It would have been more helpful if he had addressed the arguments as I actually presented them.

3.12 Returning to the argument (A), we now see how to avoid the inherent difficulties that occur with a belief system with an unrestricted domain. A sufficient thing to do is to make sure that the word "sound" is interpreted in the restricted sense which applies only to P-sentences - as was indeed done in *Shadows*, Section 3.16. (Recall the discussion of Section 2, above, in which Feferman draws attention to possible differences of interpretation of that word.) This provides the needed argument against computationalism, and it is not subject to the objection brought forward by Chalmers in his discussion of my "second argument" in his Section 2.

3.13 Of course, as in Section 7.9 of *Shadows* and as in McCullough's discussion, it is possible to repeat this argument at a higher level. Rather than restricting attention to P-sentences (that is, PI_1 sentences), we could use PI_2-sentences, say (cf. Feferman's commentary). The diagonal process can be applied, but it does not yield a PI_2-sentence, so contradiction (of the Chalmers/McCullough type - to a self-believing belief system) is again avoided. The same argument applies to higher-order sentences. However, it is important to put *some* restriction on the type of sentence to which the belief system is applied. This kind of thing is very familiar in mathematical logic. One may reason about sets, and about sets of sets, and sets of sets of sets, etc., but one cannot reliably reason about the set of *all* sets. That leads immediately to a contradiction, as Cantor and Russell pointed out long ago. Likewise, a self-believing belief system cannot consistently operate if it is allowed to apply itself to unrestricted mathematical systems. In Section 3.24 of *Shadows*, I tried to explore the tantalizing closeness that my Gödelian reasoning of Section 3.16 seemed at first to have with the Russell-type reasoning that leads to paradox. My conclusion was that the argument of Section 3.16, as I presented it, was not actually of the same nature at all, since the domain of consideration (P-sentences) was indeed sufficiently restricted. I am well aware that the argument can be taken much further than this, and it would be interesting to know how far. Moreover, it would be interesting to have a professional logician's commentary on all these lines of thinking.

4. The "Bare" Gödelian Case

4.1 Although I have concentrated, in the previous section, on what I have referred to as the "central new argument" of *Shadows*, I do not regard that as the "real" Gödelian reason for disbelieving that computationalism could ever provide an explanation for the mind - or even for the behaviour of a conscious brain.

4.2 Perhaps a little bit of personal history on this point would not be amiss. I first heard about the details of Gödel's theorem as part of a course on mathematical logic (from which I also learned about Turing machines) given by the Cambridge logician Steen. As far as I can recall, I was in my first year as a graduate student (studying algebraic geometry) at Cambridge University in 1952/53, and was merely sitting in on the course as a matter of general education (as I did with courses in quantum mechanics by Dirac and general relativity by Bondi). I had vaguely heard of Gödel's theorem prior to that time, and had been a little unsettled by my impressions of it. My viewpoint before that would probably have been rather close to what we now call "strong AI". However, I had been disturbed by the possibility that there might be true mathematical propositions that were in principle inaccessible to human reason. Upon learning the true form of Gödel's theorem (in the way that Steen presented it), I was enormously gratified to hear that it asserted no such thing; for it established, instead, that the powers of human reason could not be limited to any accepted preassigned system of formalized rules. What Gödel showed was how to transcend any such system of rules, so long as those rules could themselves be trusted.

4.3 In addition to that, there was a definite close relationship between the notion of a formal system and Turing's notion of effective computability. This was sufficient for me. Clearly, human thought and human understanding must be something beyond computation. Nevertheless, I remained a strong believer in scientific method and scientific realism. I must have found some reconciliation at the time which was close to my present views - in spirit if not in detail.

4.4 My reason for presenting this bit of personal history is that I wanted to demonstrate that even the "weak" form of the Gödel argument was already strong enough to turn at least one strong-AI supporter away from computationalism. It was not a question of looking for support for a previously held "mystical" standpoint. (You could not have asked for a more rationalistic atheistic anti-mystic than myself at that time!) But the very force of Gödel's logic was sufficient to turn me from the computational standpoint with regard not only to human mentality, but also to the very workings of the physical universe.

CHAPTER 244

4.5 The many arguments that computationalists and other people have presented for wriggling around Gödel's original argument have become known to me only comparatively recently: perhaps we act and perceive according to an unknowable algorithm; perhaps our mathematical understanding is intrinsically unsound; perhaps we could know the algorithms according to which we understand mathematics, but are incapable of knowing the actual roles that these algorithms play. All right, these are logical possibilities. But are they really plausible explanations?

4.6 For those who are wedded to computationalism, explanations of this nature may indeed seem plausible. But why *should* we be wedded to computationalism? I do not know why so many people seem to be. Yet, some apparently hold to such a view with almost religious fervour. (Indeed, they may often resort to unreasonable rudeness when they feel this position to be threatened!) Perhaps computationalism can indeed explain the facts of human mentality - but perhaps it cannot. It is a matter for dispassionate discussion, and certainly not for abuse!

4.7 I find it curious, also, that even those who argue dispassionately may take for granted that computationalism in some form - at least for the workings of the objective physical universe - *has* to be correct. Accordingly, any argument which seems to show otherwise *must* have a "flaw" in it. Even Chalmers, in his carefully reasoned commentary, seeks out "the deepest flaw in the Gödelian arguments". There seems to be the presumption that whatever form of the argument is presented, it just *has* to be flawed. Very few people seem to take seriously the slightest possibility that the argument might perhaps, at root, be correct! This I certainly find puzzling.

4.8 Nevertheless, I know of many who (like myself) do find the simple "bare" form of the Gödelian argument to be very persuasive. To such people, the long and sometimes tortuous arguments that I provided in *Shadows* may not add much to the case - in fact, some have told me that they think that they effectively weaken it! It might seem that if I need to go to lengths such as that, the case must surely be a flimsy one. (Even Feferman, from his own particular non-computational standpoint, argues that my diligent efforts may be "largely wasted".) Yet, I would claim that some progress has been made. I am struck by the fact that none of the present commentators has chosen to dispute my conclusion G (in *Shadows*, p.76) that "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth". I doubt that any will admit to being persuaded by any of the replies to my queries Q1, ..., Q20, in Section 2.6 and Section 2.10, but it should be remarked that many of these queries represented precisely the kinds of misunderstandings and objections that people had raised against my earlier use of the bare Gödelian argument (and its conclusion G) in *The Emperor's New Mind*, particularly in the many commentaries on that book in *Behavioral and Brain Sciences* (and, in particular, one by McDermott 1990). Perhaps some progress has been made after all!

5. Gödel's "Theorem-Proving Machine"

5.1 Before addressing the important issue of possible errors in human reasoning or the possible "unknowability" of the putative algorithm underlying human mathematical reasoning (which provide the counter-arguments that so many computationalists pin faith on), I should briefly refer to the discussion of Section 3.3 in *Shadows*, which Chalmers regards as "one of the least convincing sections in the book". This is the first of the two arguments of mine that he addresses, but I am not sure that he (or any other of the commentators) has appreciated what I was trying to express. In that section (and also Section 3.8, cf. figure 3.1 on p. 148), I was attempting to show the actual absurdity of the possibility that human understanding (with regard to P-sentences, say) might be encapsulated in what I have referred to as a "Gödel's theorem-proving machine". As quoted on p. 128 of *Shadows*, Gödel seemed not to have been able to rule out the possibility that mathematical understanding might be encapsulated in terms of the action of an algorithm - his "theorem-proving machine" - which, although sound, could not be humanly (unassailably) perceived to be sound. Yet it might be possible to come across this algorithm empirically. I shall refer to this putative "machine" (or algorithm) here as T.

5.2 In Section 3.3, I was concerned with a mathematical algorithm, of the type that might be considered seriously by logicians or mathematicians, so it is not unreasonable to think of T as formulated in the kind of terms which mathematical logicians are familiar with. Of course, even if T were not initially formulated in such terms, it could be if desired. It is sufficient to restrict Gödel's hypothetical theorem-proving machine to be concerned only with P-sentences. Then T would be an algorithmic procedure that generates precisely all the true P-sentences that are perceivably true, in principle, by human mathematicians. Gödel argues that although T might be empirically discoverable, the perception of its soundness would be beyond the powers of human insight. In Sections 3.3 and 3.8, I merely try to make the case that the existence of T is a very far-fetched possibility indeed, especially if we try to imagine how it might have come about (either by natural selection or by deliberate AI construction). But I did not argue that it was an entirely illogical possibility.

5.3 In Feferman's commentary, he refers to Boolos's "cautious" interpretation of the implications of Gödel's theorem that a let-out for computationalism would be the existence of "absolutely unsolvable diophantine problems". Such an absolutely unsolvable problem could be constructed, by well understood procedures, from the algorithm T, if T were to exist. Phrased in these terms, it does not seem at all out of the question that such a T might exist. In Section 3.3, my intention was merely to point out some of the improbable-sounding implications of the existence of T. It seems to me that this does go somewhat beyond what Feferman refers to at the end of his commentary. Moreover, the arguments referred to in Section 2 above (concerning Section 3.16 of *Shadows* that most commentators appear to have missed) certainly do proceed well beyond this interpretation.

5.4 Later in *Shadows* (cf. Sections 3.5-3.23, and especially 3.8), I argue that it is extremely hard to see how an extraordinarily sophisticated algorithm of the nature of T could come about by natural selection (or by deliberate AI construction), even if it could exist. It has to be already capable of correctly dealing with subtle mathematical issues that are, for example, far beyond the capabilities of the Zermelo-Fraenkel axiom system ZF (for example, the Gödel procedure can be applied to ZF to obtain humanly accessible P-sentences that are indeed beyond the scope of ZF). Yet issues of this nature played no role in the selective processes that were operative with our remote ancestors. I would argue that there is nothing wrong with natural selection having been the driving force, so long as it is a non-specific *non-computational* quality such as "understanding" that natural selection has favoured, rather than some improbable algorithm, such as T.<3>

5.5 Even if we do not worry about how T might possibly have come about, there is a distinct implausibility in its very existence, if T were to be an algorithm that could be humanly understood (or "knowable", in the terminology of *Shadows*). This is basically "case II" of *Shadows* (cf. p. 131), where the soundness of T, and certainly its specific role, would not be humanly knowable. The implausibility of such a T was the main point that I was trying to make in Section 3.3. I think Chalmers is arguing that such a T might come about by some bottom-up AI procedures and, if so, it might not look at all like a mathematical formal system. However, in the absence of some strongly held computationalist belief - to the effect that it *must* have been by procedures of this very kind that Nature was able to produce human mathematicians - there is no good reason to expect that this would be a good way of finding such a T (as I argue in *Shadows* Section 3.27), nor is there any reason to expect such a T actually to exist. It was the burden of later sections of *Shadows*, not of Section 3.3, to argue that such bottom-up procedures do not do what is required either. In effect, in these later sections, I argue that if merely the (partly bottom-up) computational *mechanisms* for ultimately leading to a T could be known, then we would indeed be able to construct the formal system that T represents. This will be discussed further in Section 7, below.

6. The Issue of Errors

6.1 Some commentators (particularly McDermott and, in effect, Baars) try to argue that the fact that human mathematicians make errors allows the computational model of the mind to escape the Gödel-type arguments. (This was also apparently Turing's let-out, as illustrated in the quote in *Shadows*, p.129.) I have stressed in many places in *Shadows* that the main arguments of that book (certainly those in Chapter 2) are concerned with what mathematicians are able to perceive *in principle*, by their methods of mathematical proof - and that these methods need not be necessarily constrained to operate within the confines of some preassigned formal system.

6.2 I fully accept that individual mathematicians can frequently make errors, as do human beings in many other activities of their lives. This is not the point. Mathematical errors are in principle correctable, and I was concerned mainly with the ideal of what can indeed be perceived in principle by mathematical understanding and insight. Most particularly, I was concerned with those P-sentences that can be humanly perceived, in principle, i.e., with those which are in principle humanly accessible. The arguments given above, in Sections 3 and 5, were also concerned with this ideal notion only. The position that I have been strongly arguing for is that this ideal notion of human mathematical understanding is something beyond computation.

6.3 Of course, individual mathematicians may well not accord at all closely with this ideal. Even the mathematical community as a whole may significantly fall short of it. We must ask whether it is conceivable that this mathematical community, or its individual members, could be entirely computational entities even though the ideal for which they strive is beyond computation. Put in this way, it may perhaps seem not unreasonable that this could be the case. However, there remains the problem of what the human mathematicians are indeed doing when they seem able to "strive for", and thereby approximate, this non-computational ideal. It is the *abstract idea* underlying a line of proof that they seem able to perceive. They then try to express these abstract notions in terms of symbols that can be written on a page. But the particular collections of symbols that ultimately appear on the pages of their notes and articles are far less important than are the ideas themselves. Often the particular symbols used are quite arbitrary. With time, both the ideas and the symbols describing them may become refined and sometimes corrected. It may not always be very easy to reconstruct the ideas from the symbols, but it is the *ideas* that the mathematicians are really concerned with. These are the basic ingredients that they employ in their search for idealized mathematical proofs. (These matters have relevance to the question of how mathematicians *actually* think,<4> as raised by Feferman in his commentary, and they are related also to issues raised also by Baars and McCullough.)

6.4 Sometimes there may be errors, but the errors are correctable. What is important is the fact is that there *is* an impersonal (ideal) standard against which the errors can be measured. Human mathematicians have capabilities for perceiving this standard and they can normally tell, given enough time and perseverance, whether their arguments are indeed correct. How is it, if they themselves are mere computational entities, that they seem to have access to these non-computational ideal concepts? Indeed, the ultimate criterion as to mathematical correctness is measured in relation to this ideal. And it is an ideal that seems to require use of their conscious minds in order for them to relate to it.

6.5 However, some AI proponents seem to argue against the very existence of such an ideal, a position that Moravec (if his robot is to be trusted as espousing Moravec's own views) seems to be taking in his commentary. Moreover, Chalmers comments: "an advocate of AI might take [the position] that our reasoning is fundamentally unsound, even in idealization". There are others, such as Baars ("I do not believe in the absolute nature of mathematical thought"), who also have difficulty with this notion, perhaps because their professional interests have more to do with examining the ways in which particular individuals may deviate from such ideals than with the ideal notions themselves. It is common for such people to point to errors that have persisted in the mathematical literature for some while (such as McDermott 's reference to Kempe's erroneous attempt at a proof of the four-colour theorem - which, incidentally provided an important ingredient in the actual proof that was finally arrived at in 1976 by Appel and Haken; cf. Devlin (1988) - or to Frege's inconsistent attempt at building up a formal set theory - which was a good deal more influential, in a very positive sense). But these errors are more in the nature of "correctable errors", and do not really argue against the very existence of a mathematical ideal.

6.6 In *Shadows*, Section 3.2, I did examine, in a serious way, the possibility that mathematical reasoning might be fundamentally unsound. But one should bear in mind that the presumption of mathematical unsoundness is an extremely dangerous position for anyone purporting to be a scientist to take. If our mathematical reasoning were indeed fundamentally unsound, then the whole edifice of scientific understanding would come crashing to the ground! For virtually all of science, at least detailed science, depends upon mathematics in one respect or another. I find it remarkable how frequently attacks on the Gödelian argument seem to degenerate into attacks upon the very basis of mathematics.<5> To attack the notion of "ideal" mathematical concepts or idealized mathematical reasoning is, indeed, to attack the very basis of mathematics. People who do so should at least pause to contemplate the implications of what they are contending.

6.7 While it is true that there are different philosophical standpoints that may be adopted by different mathematicians, this has little effect on the basic Gödelian argument, especially if we restrict attention to P-sentences; see responses to queries Q9-Q13 in Sections 3.6, 3.10 of *Shadows*. For the remainder of my arguments here, I shall take it as read that there *is* an ideal notion of (in principle) humanly accessible mathematical proof, at least with respect to P-sentences, and that this ideal notion of proof is *sound*. (And I am not against there being *more* than one, provided that they are not in contradiction with one another with regard to P-sentences; see *Shadows* Section 3.10, response to Q11.) The question, then, is how serious are the errors which undoubtedly occur when actual human mathematicians attempt toemulate this ideal.

6.8 For the arguments of Chapter 3 of *Shadows*, particularly Sections 3.4, 3.17, 3.19, 3.20, and 3.21, I try to address the issue of errors in purported mathematical arguments, and the question of constructing an error-free formal system from the *actual* output of a manifestly computational system - the hypothetical mathematical robots that I consider for the purpose. The arguments are quite intricate in places, and I do not blame some of the commentators for balking at those sections. On the other hand, it would have been helpful to have had a dispassionate discussion of these arguments in their essential points. McDermott does at least address some of the more technical arguments concerning errors - though I feel it is not altogether appropriate to refer to his account as "dispassionate". More importantly, he does not answer the essential point of my conclusions. If it is to be *errors* that provide the key escape route from the Gödel conundrum, we need to explain the seeming necessity for a "conspiracy" preventing any kind of computational procedure for weeding out all the errors in the merely *finite* set that arises in accordance of the discussion of Section 3.20 (see 3.21 and also the second paragraph of 3.28). In his commentary McDermott does not actually address the argument as I gave it, but goes off on a tangent (about a "computerized Gauss" and the like) which has very little to do with the specific argument provided in *Shadows*. (The same applies to most of his other arguments which, he contends, have "torn [my] argument to shreds". His discussion might have been more convincing had it referred to my actual arguments! I shall make some further comments concerning these matters in Section 7 below.)

6.9 McDermott does, however, come close to expressing the central dilemma presented by the Gödelian insight - although apparently unwittingly. He has a hard time coming to terms with the fact that mathematical unassailability needs "to be both informal and guaranteed accurate". Although he is unable to "see how that's possible", it is basically this conflict that forces us into a non-computational viewpoint. If by a "guaranteed accurate" notion of unassailability he means something that has been validated by a procedure that is computationally checkable, then this notion would basically have to be one that can indeed be encompassed by a formal system in the ordinary sense. We must bear in mind that the guarantee must apply not only to the correctness of *carrying out* the rules of the procedure (which is where the "computational checkability" of the procedure might have importance), but also to the validity, or *soundness* of the very rules themselves. But if we can guarantee that the rules are sound, we can also guarantee something beyond those rules. The rules would be subject to Gödel's theorem, so there would also be certain P-sentences, such as the Gödel sentence asserting the consistency of the "guaranteeing system", that would be just as "guaranteed" as the things that have already been previously "guaranteed". If McDermott is requiring that "formal" implies "computational", and that "guaranteeable" also implies computational, then he has a problem encompassing certain things that mathematicians are actually capable of guaranteeing, namely the passing from a given guaranteeing system to the implied guaranteeing of its Gödel sentence.

6.10 One of the key points of the discussion of Chapter 3 of *Shadows* was to exhibit the importance of this conflict within the context of an entirely computational framework. If we accept that the putative robots described there are entirely computational entities, then any "guaranteeing" system that they come up with must necessarily be computational also. Accepting that the robots must also guarantee their guaranteeing system (see Section 3 above) and that they appreciate Gödel's theorem - and also accepting that random elements play no fundamentally important role in their behaviour (see 3.18, 3.22) - we are driven to the remaining loophole for computationalism: errors. It was the thrust of Sections 3.17-3.21 to demonstrate the implausibility of this loophole also. For this discussion, one attempts to find computationally bounded safeguards against errors, and then shows that this is impossible.

6.11 In effect, though in a stronger form than usual, all this is saying is that it is impossible to "formalize" the informal notion of unassailable mathematical demonstration. In this sense McDermott is indeed right to fail to "see how that's possible". It's *not* possible if "formalize" indeed implies something *computational*. That's the whole point!

7. The "Unknowability" Issue

7.1 Several other commentators (Chalmers, Maudlin, Moravec - and also McDermott again!) prefer to attack the Gödel argument from the standpoint that the "algorithm" (or formal system) to which Gödel's theorem is to be applied is *unknowable* in some sense - or, at least, unknowable to the person attempting to apply the argument. (Indeed, Chalmers, for one, seems to be happy enough to accept "that we have an underlying sound competence, even if our performance sometimes goes astray"; so in his commentary on my "First Argument" - that given in *Shadows*, Section 3.3 - he seems to be resorting to the "unknowability" of the algorithm in question.)

7.2 There is an unfortunate tendency for some people (Chalmers, and some others excepted) to try to twist my use the Gödel argument away from the form in which I actually gave it, which refers to "mathematical understanding" in the abstract sense - or at least in the sense in which that term might apply to the mathematical community as a whole - to a more personal form. Such people seem to regard it as more impressively ridiculous that some individual mathematician could know his or her "personal algorithm", than that the principles underlying the proof procedures that are common to mathematicians as a whole might be accessible to the common understanding of the mathematical community. And they apparently regard it as particularly evidently ridiculous that I myself should have such access (cf. commentaries by McCullough, Maudlin, and Moravec), so they phrase what they take to be my own Gödelian arguments in the form of what kind of a contradiction I might land myself in if I happened to come across my own personal algorithm! I suppose that in order to make "debating points", such procedures may seem effective, but I find it distinctly unhelpful to phrase the arguments in this way; for the arguments then become significantly changed from the ones that I actually put forward.

7.3 Particularly unhelpful are formulations like Moravec's "Penrose must err to believe this sentence." and McCullough's "This sentence is not an unassailable belief of Roger Penrose." Although there are ways of appreciating the nature of the particular sentence that Gödel originally put forward in terms that are not totally dissimilar from this, it is certainly a travesty to attempt to express the essentials of my own (or indeed Gödel's) argument in this way. Only marginally better would be "No mathematician can believe unassailably that this sentence is true." or "No conscious being can accept the truth of this sentence." - mainly because of their manifest similarity to the archetypal self-contradictory assertion: "This sentence is false." In Section 3.24 of *Shadows*, I explicitly addressed the possibility that the kind of reasoning that I had been using earlier in the book (basically the argument of 3.16, which is that of Section 3 above, but also 3.14) might be intrinsically self-contradictory in this kind of way. I do not think that it is, for reasons that I discussed in 3.24. None of the commentators has chosen to dispute me on this particular issue, so perhaps I may take it that they agree also!

7.4 Instead, the arguments, relevant to the present discussion, that Chalmers, Maudlin, McDermott, and Moravec are really putting forward (and which are greatly obscured by the above kind of formulation), is that the algorithm in question might be *unknowable*. They make the point that in order to provide an effective simulation of the thought processes of an individual mathematician, an almost unimaginably complicated algorithm would have to be envisaged. Of course, this point had not escaped me either(!), which is the main reason why I formulated my own discussion in quite different ways from this.

7.5 There are, in fact, two distinct broad lines of argument put forward in *Shadows*, the *simple* argument and the *complicated* argument. The simple argument (which has always been good enough for me) is basically the "bare" Gödelian reasoning referred to in Section 4 above (leading to the conclusion G of *Shadows*, p.76), as applied to the mathematicians' belief that they are "really doing what they think they are doing", rather than blindly following the rules of some unfathomable algorithm (see the opening discussion of 3.1 and the final one of 3.8). Accordingly, the procedures available to mathematicians *ought* all to be knowable! The only remark concerning any aspect of implications of this line of argument that I can find in these commentaries is that towards the end of McDermott's piece, in which he remarks that the quality of conscious understanding will, in his view, turn out to be something "quite simple" (because "consciousness is no big deal"). I remarked (*Shadows*, p.150) that "understanding has the appearance of being a simple and common-sense quality", but if it actually is something simple, it has to be something non-computational, because otherwise it becomes subject to the bare form of the Gödelian argument. I do not think that McDermott would be very happy with that, but he does not refer to this particular problem. (As an aside, I find it hard to see why some commentators, such as Maudlin, seem to argue that the slightest flaw in the discussion of Part 1 of *Shadows* would demolish the whole argument. In fact there are several different lines of argument presented there. All of them would have to be demolished independently!)

7.6 The complicated lines of argument are addressed more at those who take the view that mathematicians are not "really doing what they think they are doing", but are acting according to some unconscious unfathomable algorithm. Since there is no way that we could know what this algorithm is (or what several distinct but effectively equivalent algorithms might be; cf. *Shadows*, Section 3.7), I adopt a completely different line of approach. This is to examine how such an unfathomable algorithm might conceivably come about. The issue of the role of natural selection (treated particularly in 3.8) was referred to in Section 5 above. The other possibility that I discussed was some form of deliberate AI construction, and that was the thrust of Sections 3.9 onwards, in Chapter 3 of *Shadows*.

7.7 Rather than trying to "know" whatever putative algorithm might now describe the physical action of the brain of some individual human mathematician - or else what complicated computer program might now control the actions of some putative intelligent mathematics-performing individual robot - I consider the general type of computational AI process that might underlie the evolution of such a robot. We do not need to know how the robot's computer-brain is actually supposed now to be wired up, since I am prepared to accept that the "bottom-up" procedures that are used (artificial neural networks, genetic algorithms, random inputs, even natural selection processes that might be applied to the robots themselves, etc. - and also adequately simulated environments, cf. *Shadows* Section 3.10, McCarthy, McCullough, and McDermott take note) could lead to a final product of almost unimaginable complication. Nevertheless, these very *mechanisms* that go into the ultimate construction of the robots would indeed be knowable - in fact, it might well be claimed (as I know that Moravec (1988) has actually claimed) that these mechanisms are, in effect, known already. The whole point of considering these mechanisms, rather than the "actual" algorithm that is supposed to be enacted by the computer-brain of our putative robot (a point apparently missed by Maudlin, McDermott, and Moravec), is that the former would be supposed to be knowable, so long as those aspects of the AI programme that are aimed at the construction of an actually intelligent robot - intelligent enough for it to be able to understand mathematics - are attainable within the general framework of present-day computer-driven ideas.

8. AI and MJC

8.1 A summary of this line of reasoning formed part of the "fantasy dialogue" given in Section 3.23 of *Shadows*. (In what follows, "MJC" refers to the robot, whereas "AI" refers to the subject of artificial intelligence.) Thus, when Maudlin ridicules the possibility that MJC might "easily 'digest' its own algorithm", he has missed the point. There is not supposed to be anything "unknowable" about the procedures of AI; otherwise there would be no point in people actually trying to *do* AI!

8.2 One of the aims of the discussion in the dialogue was to bring home the fact that, according to the "optimistic" school of AI, to which Moravec belongs, it need not be so far into the future when robots are actually constructed which could exceed all human mental capabilities. In particular, such a robot could perform feats of mathematical understanding that exceed those of any human mathematician. This indeed seems to be a corollary of such an optimistic stance with regard to AI, and is not particularly (as McDermott contends) "extravagant" from the point of view of the tenets of AI. My characterization of MJC was to set its abilities, with regard to mathematical understanding, just ahead of that of humanity, but with a particularly effective ability with regard to directly computational matters. Thus, it would have no difficulty at all in assimilating the purely computational aspects of the mechanisms concerning its original construction (since these were in any case already known to Albert Imperator, but they might be computationally very involved), whilst MJC might be relatively slower in appreciating the subtleties of certain logical points - although still a good deal faster than one might imagine a human mathematician would be.

8.3 This does not seem to me to be an "incongruity" in characterization, as Moravec seems to suggest. Of course MJC goes mad at the end - but why not? It has just been driven to the logical conclusion that the only way in which it could have come about was by God implanting a Divine Algorithm into its mechanisms, through the "chance" elements that were part of those mechanisms. It is not a question of MJC suddenly realizing that its initials stand for "Messiah Jesus Christ", as McDermott seems to think. (The initials were just intended as a little joke for the reader, and not really part of the story.) In fact, McDermott seems extraordinarily slow in getting to the point of the story, if indeed he ever really gets to the point of it. (Actually, it seems that he does not, especially in view of his comments about "affixing a * to Omega(Q)", etc. He has not appreciated the central argument repeated in Section 3 above. It is clear that Maudlin misses the point also, since the dialogue has nothing to do with "a computer failing to pass the Turing test". But so also does Moravec's robot, so McDermott and Maudlin are in good company!) I certainly do *not* believe that a computationally controlled robot could achieve the kind of easy-flowing intelligent-sounding dialogue that MJC exhibits. That is the whole point of a reductio ad absurdum. One assumes that all the implications of the premise, that one intends ultimately to disprove, actually hold good. The final contradiction disproves the premise. Here, the premise is that the procedures of computational AI can ultimately lead to the construction of an intelligent mathematics-performing robot. Of course such a putative robot could be articulate and sound intelligent in other ways than just in mathematics. But it doesn't mean that I believe the premise.

8.4 On another point, the fantasy dialogue does not actually summarize *all* the arguments of Chapter 3 of *Shadows*. Most particularly, it does not summarize most of the arguments given in Sections 3.17-3.21 against the "errors" argument (cf. Section 6 above). I did not include these mainly because I felt that the discussion was already getting rather long and complicated; and since the "error" discussion was rather involved, I thought it best to leave most of it out. In addition to this, the way that the dialogue developed, it seemed appropriate for MJC to have a distinctly arrogant character. It would have changed the flavour of the story to allow MJC to acquire the humility that would have been needed in order to have it admit to being subject to serious error.

8.5 In some ways this was perhaps unfortunate, because it appears to give Maudlin and McDermott an easy way out by allowing their allegedly more "realistic" version of MJC to make the occasional mistake. This, however, would be to miss the point of the "errors" arguments, as given in *Shadows* and as referred to in Section 6 above.

9. Mathematical Platonism

9.1 I think that a few remarks in relation to my attitude to mathematical Platonism are appropriate at this stage. Indeed, certain aspects of my discussion of errors, as given in Section 6 above, might seem to some to be inappropriately "Platonistic", as they refer to idealized mathematical arguments as though they have some kind of existence independently of the thoughts of any particular mathematician. However, it is difficult to see how to discuss abstract concepts in any other way. Mathematical proofs are concerned with abstract ideas - ideas which can be conveyed from one person to another, and which are not specific to any one individual. All that I require is that it should make sense to speak of such "ideas" as real things (though not in themselves material things), independent of any particular concrete realization that some individual might happen to find convenient for them. This need not presuppose any very strong commitment to a "Platonistic" type of philosophy.

9.2 Moreover, in the particular Gödelian arguments that are needed for Part 1 of *Shadows*, there is no need to consider as "unassailable", any mathematical proposition other than a P-sentence (or perhaps the negation of such a sentence). Even in the very weakest form of Platonism, the truth or falsity of P-sentences is an absolute matter. I should be surprised if even Moravec's robot could make much of a case for alternative attitudes with regard to P-sentences (though it is true that some strong intuitionists have troubles with unproved P-sentences). There is no problem of the type that Feferman is referring to, when he brings up the matter of whether, for example, Paul Cohen is or is not a Platonist. The issues that might raise doubts in the minds of people like Cohen - or Gödel, or Feferman, or myself, for that matter - have to do with questions as to the absolute nature of the truth of mathematical assertions which refer to large infinite sets. Such sets may be nebulously defined or have some other questionable aspect in relation to them. It is not very important to any of the arguments that are given in *Shadows* whether very large infinite sets of this nature actually exist or whether they do not or whether or not it is a conventional matter whether they exist or not. Feferman seems to be suggesting that the type of Platonism that I claimed for Cohen (or Gödel) would require that for *no* such set could its existence be a conventional matter. I am certainly not claiming that - at least my own form of Platonism does not demand that I need necessarily go to such extremes. (Incidentally, I was speaking to someone recently, who knows Cohen, and he told me that he would certainly describe him as a Platonist. I am not sure where that, in itself, would leave us; but it is my direct personal impression that the considerable majority of working mathematicians are at least "weak" Platonists - which is quite enough. I should also refer Feferman to the informal survey of mathematicians reported on by Davis and Hersch in their book The Mathematical Experience, 1982, which confirms this impression.)

9.3 The issue as to the "existence" of some very large set might occasionally have a bearing on the truth or otherwise of certain P-sentences. Accordingly, a mathematician's belief with regard to such a P-sentence might be influenced by that mathematician's particular attitude to the existence of such a set. Questions of this nature were discussed in *Shadows*, Section 2.10, response to Q11, where it is concluded that there is no great issue to disturb significantly the Gödelian conclusion G. Feferman has not chosen to comment on this matter, so I suppose that he has no strong objection to my line ofreasoning.

10. What has Gödel's Theorem to do with Physics?

10.1 Maudlin questions the very basis of my contention that one can indeed deduce something important and new about the nature of physical laws from the actual behaviour of certain physical objects: human mathematicians.<6> As far as I can make out, his basic claim is that the computability, or otherwise, of mathematicians has no externally observable consequences. I find this claim to be a very strange one. He refers to what he calls the "Strong Argument", which he says is "clearly unsound". The Strong Argument contends that "no computer could reliably produce the visible outward motions of a conscious person" and, consequently, there must be something beyond computation in the behaviour of physical objects (e.g. humans). Maudlin's objection seems to rest on the *finiteness* of the total output of a human being. Whatever the total output of some human being might be (and his "human being" is "Penrose", of course!), that output would indeed be finite. Therefore there would be some computer program which could, in principle at least, simulate that person's action.<7> This is a very odd line of reasoning, because it would invalidate any form of deduction about physical theory from observation whatsoever. The number of data points concerning observations of the solar system is finite, after all, so those data points could form the output of a sufficiently large computer, quite independently of any underlying physical theory. (Or they could be used to support a wrong theory with enough parameter freedom, such as the Ptolemaic theory, or even chariots in the sky.) I am tempted to reply to Maudlin by merely saying "be reasonable!"

10.2 Of course, canned answers could in principle provide any answer you want - even with infinite numbers of alternatives if the canning is allowed to be infinite. But the *whole point* of a Turing test (as Turing himself importantly understood) is that it takes the form of a question and answer session. It is simply not *practicable* to take into account all conceivable questions and follow-up questions and follow-up follow-up questions, etc. simply by storing all possible alternatives. (Anyone who has contemplated the task of writing a comprehensive CD-ROM program - or even a book such as *Shadows* in which one attempts to "second guess" all readers' possible counter-arguments - will begin to appreciate what I mean. There can be a significant complexity explosion even in the relatively short reasoning chains that are involved in such things.) Maudlin refers to this matter of complexity explosion, but he does not draw the appropriate conclusion from it.

10.3 My contention is that without any *genuine understanding* on the part of the computer, it will (at least in most cases) eventually be found out, when subjected to sensitive enough questioning. Trying to simulate intelligent responses by having mountains and mountains of stored-up information, using the programmer's best attempts to assimilate all possible alternatives, would be hopelessly inefficient. It appears that Maudlin believes that he has made a decisive logical ("in principle") point by bringing in the finiteness argument. But he is allowing his computer to have an exponentially larger finite storage limit than the finite limit that he imposes on the human (which is a general feature of the "canned response" approach), and this is totally unreasonable. Indeed, this "exponential" relationship involved in a canned response (or in what is called a "look-up table") *is* a decisive logical ("in principle") response to Maudlin's proposal. This applies both in the finite and in the (idealized) infinite case; for we have 2^alpha > alpha whether or not alpha is finite, and this inequality comes from the same kind of diagonal argument (Cantor's original argument) as that used in the Gödel theorem.

10.4 In fact the finiteness issue was discussed in *Shadows* (in the responses to Q7 and Q8 in Section 2.6), though from a slightly different angle. Maudlin does refer to this discussion, but he appears to misunderstand it. (Baars, in expressing his somewhat muddled parallel/serial worries about the "infinite memory" of a Turing machine is in effect, also addressing the "finiteness" issue, but he does not refer to my discussion of it, nor to the relevant Section 1.5.) In that discussion, I addressed the problem of how one might provide answers to mathematical questions - of, say, deciding the truth of P-sentences - by simply listing all the correct answers. In my response to Q7, I pointed out that the very process of listing the answers required some means of forming reliable truth judgements. This matter has simply been ignored by Maudlin, yet it contains the whole point of the non-computability argument. In order to be able to list the *correct* answers to the P-sentences in his canned responses, Maudlin's *computer programmer* will need to possess the (non-computable) quality of understanding in order to provide what are actually the correct answers! When I said, in *Shadows* that "the odd against this are absurdly enormous", I was referring to the chances against providing the answers to mathematical problems of this nature without any understanding on the part of the programmer. Maudlin's situation is completely different, where he in effect presupposes that the programmer is allowed to have this understanding, and this completely begs the non-computability question.

10.5 There is, however, a somewhat related issue that has also been raised with me by other people: how could one actually *tell*, by observational means alone, whether or not the physical world behaves non-computably? (Here, I am leaving aside the question of the behaviour of extremely highly sophisticated physical objects like human beings; I am concerned with direct physical experiments and the like.) It seems to me that this issue is quite comparable to a somewhat related one, namely that of *determinism*. How could one tell by direct physical experiment whether or not the physical world is deterministic? Of course, one cannot tell - not just like that. Yet there is the common assertion that the classical behaviour of physical objects is indeed deterministic. What this means is that Newtonian *theory* (or Maxwell's *theory* or Einstein's *theory*) is deterministic; that can be shown mathematically. What one does is to design sophisticated experiments or observations to test the theory in other respects, and if the expectations of the theory are borne out, we conclude that various other things about that theory, such as the fact that it is indeed deterministic, ought also to hold for the behaviour of the universe (to the appropriate degree of approximation as is implied by the limits within which the theory has been shown to be valid). And so it will be with the new theory of physics that unites the classical and quantum levels and which, I maintain, will turn out to be a non-computable theory. Of course, I am at a disadvantage here, since this theory has yet to be discovered! But the general point is the same.

11. How Could Physics Actually Help?

11.1 Several commentators (Baars, Chalmers, Feferman, Maudlin) question the competence of *any* physical theory ever having anything of importance to say about mind, consciousness, qualia, etc. and Klein asks for clarification on this issue. According to Feferman, for example, my attempts to push the consciousness discussion in the direction of physics would merely be to replace one "nothing but" theory with another, i.e. to replace "the conscious mind is nothing but a computer" with "the conscious mind is nothing but sub-atomic physics". Other commentators, in effect, express similar worries. In fact, to describe things in the aforementioned way is rather to miss the point of what I am trying to say. I certainly do not expect to find any answers in sub-atomic physics, for example. What I am arguing for is a *radical upheaval* in the very basis of physical theory.

11.2 In most respects, this upheaval would have to have no observable effects, however. This might seem odd, but we have an important precedent. Einstein's general relativity, as regards most (indeed, almost all) of its observational consequences, is identical with Newton's theory of gravity. Yet, it indeed provided a radical upheaval in the very basis of physical theory. The concept of gravitational force is gone. the concept of a flat background Euclidean space is gone. The very fabric of space-time is warped, and the density of energy and momentum, in whatever form, directly influences the measure of this warping. The precise way in which the warping occurs describes gravity and tells us how matter is to move under its influence. Self-propagating ripples in this space-time fabric can occur, and carry away energy in a mysterious non-local way. Although for many years observational support for Einstein's theory was rather marginal, it can now be said that, in a clear-cut sense, Einstein's theory is confirmed to a precision of one part in one hundred million million - better than any other physical theory (see *Shadows*, Section 4.5).

11.3 What I am asking for is a revolution of (at least) similar proportions. It should represent as much of a change in our present-day ways of looking at quantum theory as general relativity represents a change from Newtonian theory. Some will argue, however, that even the profound changes that I have described above, which overturn the very basis of Newtonian physics, will do nothing to help us come to terms with the puzzle of mentality within such a physically determined universe. I do not deny the significance of that argument. But we do not yet know the very form that this new theory must take. It might have a character so different from that which we have become accustomed to in physical theory that mentality itself may not seem so remote from its form and structure. Moreover, quite apart from any considerations of mentality, there are, in my opinion, very powerful reasons coming from within physics itself for believing that such a revolution is necessary. (Baars, in particular, fails to appreciate this point when he says "there is yet nothing to revolt against".)

11.4 Einstein's theory was to do with the issue of how to describe the phenomenon of gravity - in its action in guiding the planets and the stars and the galaxies, and in the shaping of the large-scale structure of the universe. These phenomena do not directly relate to the processes which control the behaviour of our brains and which presumably actually underlie our mentality. What I am now asking for is a revolution that would operate at the very scales relevant to mental processes. Yet I am also arguing that the physical revolution we seek should actually be dependent upon the particular revolutionary changes that Einstein's theory already represented from the older Newtonian ideas about the nature of reality.

11.5 I know that this puzzles many people; in fact, it puzzles many *physicists* that I should seriously attempt to claim such a thing. For the scales at which gravitational interactions reign seem totally different from those which operate in the brain. A few words of explanation may well be helpful at this juncture. I am certainly not asking that gravitational interactions (or "forces") should have any significance for the physical processes that are going on in the brain. The point is quite a different one. I am referring, instead, to the influences that Einstein's viewpoint with regard to gravity will have upon the very structure of quantum theory. Instead of quantum superpositions persisting for all time - as standard quantum theory would have us believe - such superpositions constitute a state which is *unstable* (see Penrose 1996). Moreover, this decay time can be computed, at least in certain very clear-cut situations. Yet, many physicists might well take the view that the time-scales, distance-scales, mass-scales, and energy-scales that would arise in any framework that purports to embody the union of Einstein's general relativity with quantum theory must be hopelessly wrong. Indeed the relevant time-scale (~10^-43 seconds) is some twenty orders of magnitude shorter that the briefest processes that are considered to take place in particle physics; the relevant space-scale (~10^-13 cm) is some twenty orders of magnitude smaller than the diameter of a proton; the relevant mass-scale (~10^-5 grams) is about the mass of a flea, which seems much too big; and the relevant energy scale (~10^18 ergs) is about what would be released in the explosion of a can of petrol. However, when one comes to examine the details, these figures conspire together (some being individually too small but others correspondingly too big) to produce an effect that is indeed of an eminently appropriate magnitude. (For details, see a forthcoming paper by Stuart Hameroff and myself (Hameroff and Penrose 1996).)

11.6 Again, many would argue that we shall still have come no closer to an understanding of mentality in physical terms. Perhaps, indeed, we will not have come a great deal closer. But I believe that some progress will have been made in an appropriate direction. The picture of quantum state reduction that this viewpoint is concerned with ("OR": objective state-reduction) involves the bifurcation and then selection of one out of several choices for the very shape of space-time. Moreover, there are fundamental issues arising here as to the nature of time and the apparent flow of time (see Section 13 below, in relation to Klein's commentary). I am not arguing that these issues will, in themselves, resolve the puzzles of human mentality. But I do claim that they could well point us in new directions of relevance to them, and this could change the very nature of the questions that the problems of mentality raise.

11.7 I think that people in AI, and perhaps a good many philosophers also, have a tendency to underestimate the importance of the *specific* nature of the physical laws that actually govern the behaviour of our universe. What reason do we really have to assume that mentality does not need these particular laws? Could consciousness arise in a world controlled by some arbitrarily chosen set of rules? Could it arise within scope of John Conway's "game of life" (Gardner 1970, Poundstone 1985), for example, as Moravec (1988) has suggested? Although the Conway rules for a "toy universe" are ingenious, they do not have the subtle sophistication of Newtonian mechanics - whose sophistication people often take for granted. Yet despite the extraordinary fruitfulness of Newtonian ideas, even they cannot explain something so basic as the nature and stability of atoms. We need quantum theory for that. And even quantum theory does not fully account for the behaviour of atoms, because its explanations require that curious hybrid of procedures of unitary (Schroedinger) evolution and quantum state-vector reduction (denoted in *Shadows* by U and R, respectively) - procedures which are not really consistent with one another, I claim.

Eventually, in order to explain even the stability and the specific nature of atoms, we shall need a better theory of physics than we have today, at the *fundamental* level.

11.8 There is no doubt that physics - and often the very detailed nature of the specific underlying physical laws - is essential to most of the sophisticated behaviour of the world we know. So why should the most sophisticated behaviour that we know of in the world, namely that of conscious living human beings, not also depend on the very detailed nature of those laws? As I have indicated above, we do not yet know the full nature of these laws, even in some of their most basic respects. A new theory is needed quite independently of any necessity for new laws to describe a universe that can support consciousness. However, physicists themselves often get carried away into thinking that they know everything that is needed - in principle, at least - for the behaviour of all things of relevance. There is a curious irony, here, in McDermott's quoting from *Shadows* p.373 "It is only the arrogance of the present age that leads so many to believe that we now know all the basic principles that can underlie all the subtleties of biological action." For he takes that remark to be aimed primarily at the AI community. In fact, the people I had primarily in mind were the (theoretical) *physicists*. I do not blame the biologists, or even AI researchers, when they take from the physicists a picture of the world commonly claimed to be almost final - bar some technical details that are irrelevant for the behaviour of macroscopic objects. But perhaps McDermott is right; some AI researchers seem to be nearly as arrogant as high-energy physicists (and with far less reason) - especially those AI researchers who claim that the deepest mystery of the physical world can be answered without any reference to the actual laws that govern that world!

11.9 I should make it clear, however, that I am certainly making no claim that the mystery of mentality can be resolved *merely* by finding the correct physical theory. I am sure that there are vital insights to be gained from psychology as well as from neuro-physiology and other aspects of biology. Baars seems to think that I am denying the existence of the *unconscious*, because there is no significant mention of it in *Shadows* (though there was some small reference to the unconscious mind in *The Emperor's New Mind*). I should like to reassure Baars that I fully accept both the existence of the unconscious and its importance to human behaviour. The only reason that the unconscious was not discussed in *Shadows* was that I had no contribution to make on the subject. I was concerned with the issue of *consciousness* directly, in particular in relation to the quality of understanding. However, I certainly agree that a complete picture cannot be obtained without the proper role of unconscious mentality being appreciated also.

12. State-Vector Reduction

12.1 Some commentators express worries in connection with my quantum state-vector proposals - whereby the quantum procedure R is to be replaced by some form of *objective* reduction, which I denote by OR. There are many misunderstandings here. Baars seems to think that I am taking the view that R has something to do with "observer paradoxes", which is explicitly not my view, as I thought I had made clear in *Shadows*, Chapter 6. Klein does not make this mistake, but seems to take the view that the measurement problem (R) has (or ought to have) something directly to do with metaphysics. This is certainly different from my own "objective" standpoint with regard to R.

12.2 Maudlin complains that my "objections to Bohm's theory" (a theory that, in a sense, incorporates R) "are impossible to decipher from the text" - which is not surprising since I did not give them there - and that my "objections to the GRW theory are clearly not decisive". My objections to GRW (the OR scheme of Ghirardi, Rimini, and Weber, 1986) were not meant to be decisive. In my opinion, this scheme is a very interesting one, but it suffers from being somewhat ad hoc. What one needs (and I am sure that the authors of this scheme would not disagree) is some way of fitting the scheme in more convincingly with the rest of physics. In fact Diosi made a proposal in 1989 that could be regarded as a GRW-type model in which the ad hoc nature of the GRW parameters was removed by fixing them to be provided by the *quantum gravity* quantities referred to in Section 11, above. Diosi's model encountered difficulties, as was pointed out by Ghirardi, Grassi, and Rimini (1990), who also suggested a remedy, but at the expense of re-introducing another parameter. It should be said that in fact the Diosi-Ghirardi-Grassi-Rimini proposal is extremely close to the OR scheme that I was proposing in *Shadows* (and in Penrose 1995). These other authors do not mention non-computability (their proposal being entirely stochastic), but there is no essential incompatibility between both sets of ideas. In Sections 7.8 and 7.10 of *Shadows*, I give some reasons (admittedly far from conclusive) for anticipating that a full quantum gravity scheme of this nature might indeed be non-computable.

12.3 In his final paragraph, Maudlin seems to be complaining that the tentative OR proposals that are being promoted in *Shadows* do not solve all the problems of uniting quantum theory with relativity, and of explaining the problems of human cognition. He's not asking for much! These proposals were hardly intended to provide a complete theory (as anyone reading Section 7.12 of *Shadows* would surely appreciate) but merely to give some idea of the orders of magnitude involved in the collapse rate for such an OR theory - if such a theory *were* to be found.

12.4 Klein refers to the excellent little book QED of Feynman (1985) which introduces the basic rules of quantum theory (and quantum fieldtheory) with the minimum of fuss. However, Feynman never attempts to address the measurement problem in this book - which amounts to the issue of why (and when) do the quantum-level complex-valued amplitudes become classical-level real-valued probabilities, in the process of having there moduli squared. It might be worth mentioning that I read QED for ideas, just before embarking on writing my chapter on quantum mechanics in *The Emperor's New Mind*. However, I found that Feynman's approach was not altogether suitable for me because I needed to address the measurement problem in some detail, which Feynman avoided completely. Feynman has certainly worried about this problem, but he preferred not to emphasize it in his writings. There is a historical point of interest, here. For it was actually Feynman's early worrying about the nature of the union of Einstein's general relativity with quantum mechanics (expressed in Feynman's contribution to the conference held in Chapel Hill in the 1950s, cf. Feynman et al 1995, Section 1.4, p. 15) that originally motivated Karolyhazy (1966) to seek an explanation of state-vector reduction in terms of gravitational effects (and Feynman also influenced me in the same way). Diosi's particular approach arose from the work of Karolyhazy's Budapest school.

12.5 Questions to do with "the overlap of states" referred to by Klein do not really resolve the measurement issue, and von Neumann's point about the difficulty of locating exactly where (or when) R takes place just emphasizes the subtlety of the R phenomenon. However Klein is completely right in pointing to the biological difficulties involved in maintaining quantum coherence within microtubules and, more seriously, in allowing this coherence to "leap the synaptic barrier". To see how this might be achieved is a fundamental problem for the type of scheme that I (in conjunction with Stuart Hameroff) have been proposing. Clearly more understanding is needed. (See Section 14 below, for a tentative suggestion in relation to this.)

12.6 I should point out a misunderstanding on the part of Maudlin. He seems to think that my "collapse theory offers a stochastic collapse postulate" and that it is concerned with "the exact timing of the collapses". I have nowhere said either of these things. I get the impression that Maudlin has been confused by the comparisons that I have been making between the suggested OR process and the phenomenon of the decay of an unstable particle (or unstable nucleus). But at no stage did I suggest that it would be in the precise *timing* of the decay of the quantum superposition that significant non-computability would occur. (But on rereading the relevant parts of *Shadows*, I realize that I was not at all specific there as to what I *did* mean.) Of course, since the detailed theory is not known, it is possible that there would be relevant non-computability in the timing also, but what I had in mind was something quite different, and certainly more relevant. The idea is that when collections of tubulin conformational movements become involved in coherent quantum superposition, there will come a point when the mass movements within the tubulin molecules are sufficient for OR to come into effect (without significant environmental disturbance). When that happens Nature must make a choice between the various collections of conformational states under superposition. It is not so much a question of *when*, but of *which* among the collection of superposed states Nature indeed chooses. Choices of this kind could actually influence the behaviour of a synapse. (There are various possibilities for this; for example, the particular collections of conformational states of tubulins in a microtubule might influence a dendritic spine, via the actin within the spine. Moreover, a great number of microtubules would be expected to act in concert - since a single OR state-selection process would act within many microtubules all at once. However, there is no point in trying to be too specific at this stage.) It would be in the *particular* choice that Nature makes that the non-computability could enter significantly, and this particular choice (global over a significantly large area of the brain, probably involving at least thousands of neurons) could result in subtle collective changes of synapse strengths all at once. (See Hameroff and Penrose 1996.)

13. Free Will

13.1 What kind of a theory might it be that determines these choices? Many people who are unhappy with computationalism would be just as unhappy with any other type of mathematical scheme for determining them. For they might argue that it is here that "free will" makes its entry, and they would be unhappy that their free-will choices could be determined by *any* kind of mathematics. My own view would be to wait and see what kind of non-computable scheme ultimately emerges. Perhaps a sophisticated enough mathematical scheme will turn out not to be so incompatible with our (feelings of) free will. However, McCarthy takes the view that I am "quite confused" about free will, and that my ideas are "not repairable". I am not really clear about which of my confused ideas McCarthy is referring to. In *Shadows*, I did not say much about the issue of free will, except to raise certain issues. Indeed, I am not at all sure what my views on the subject actually are. Perhaps that means that I *am* confused, but I do not see that these ideas are remotely well enough defined to be irreparable!

13.2 As I remarked above, most people would probably take the view that if there is *any* kind of mathematical theory precisely determining the way we behave, then there is no free will. But, as I have indicated, I am not so sure about this. The answer could depend on the very nature of this mathematical theory. The theory would certainly have to be non-computable (according to my own considerations), but much more than this. We recall from the discussion given in Section 3 above (McCullough, Chalmers, and Section 7.9 of *Shadows*) that the Gödel diagonalization procedure can be applied to systems much more general than merely computational ones. Thus, my arguments would equally imply that our missing theory must be not just non-computational, but also beyond (or at least different from) Turing's notion of oracle computation. (An oracle computation is what is achieved by a Turing machine to which an additional command is appended: "if the pth Turing machine acting on the number q eventually halts, do A; if it doesn't, do B".) Again, we can consider second-order oracle machines (which can assess whether first-order oracle machines ever halt), and the diagonalization still applies. So the missing theory is not a second-order oracle theory either. The same applies to even higher-order oracles. Indeed the missing theory cannot be an oracle theory of order alpha for any computable ordinal number alpha. As far as I can make out, this is not the limit of it either. Diagonalization can be applied in very general circumstances indeed. We enter very nebulous areas of mathematical logic. It seems that the quality of "understanding" - which is what this discussion is effectively all about - is something very mysterious. Consequently, any theory of the physical world which is capable of accommodating beings that are capable of genuine understanding must itself be in a position to cope with such subtleties.

13.3 As a side comment, I should remark that this form of "repeated Gödelization" is somewhat related to, but not at all the same as, that referred to by McCarthy and McCullough, who both describe the process whereby sound extensions of a sound formal system can be obtained, corresponding to any computable ordinal alpha. This procedure was described in *Shadows* Section 2.10, answer to query Q19. I am not quite sure why they go to the trouble to repeat this argument, with no reference to my own discussion. The conclusion, noted in *Shadows*, that "repeated Gödelization does not provide us with a mechanical procedure for establishing the truth of P-sentences" is confirmed by Feferman. (As far as I can make out, Feferman's comments about how his work extended that of Turing's are related to the considerations of the previous paragraph, above.)

13.4 The issue of free will is related also to the experiments of Libet (1992), (and to earlier experiments of Deeke, Groetzinger, and Kornhuber (1976) and also Grey Walter) that are referred to by Klein. These experiments suggest a delay of the order of a second, in an entirely volitional act, between the first indications of mental activity (as evidenced by brainwave studies) and the final willed (say) finger movement. Klein calls into question my puzzlements, expressed in *Shadows* Section 7.11, concerning the seeming slowness of consciousness - but as far as I can make out, he has misunderstood my point (as, I believe, did Ian Glynn before him, in the

1990 article that Klein refers to). Klein says that there are "no surprises" in the fact that there is "substantial unconscious processing" (in fact, of the order of a second's worth) before "the subjective awareness of the decision to act" takes place, and that only about one fifth of a second's delay occurs before the motor act. But it was *precisely* the length of time involved in the unconscious processing that was worrying me. (Or at least this was *one* of the two things that was worrying me; the other had to do with the *passive* half-second delay that Klein also regards as an over-estimate - to which I shall return below.)

13.5 If consciousness has any active role to play in a response to an external stimulus, then it is no good for the unconscious to have already lined up the action a second ahead of time unless the unconscious already "knew" what decision that the conscious mind was going to take. Klein asserts that I am referring to a "stimulus-response situation", for which the response could be much more rapid and essentially entirely *unconscious*. In fact I am *not* considering situations of this kind, but those in which it is necessary for consciousness to come into play in order that it can actively influence the outcome. If the "free will" of the conscious mind is allowed to come into play, then surely it would be necessary for *all* processes that need to be involved, whether they be the conscious ones or the preparatory unconscious activities, to take place *after* the external stimulus occurs. This leads us to a about a second's delay for a consciously influenced response.

13.6 The extra half second comes from Libet's other (passive) experiment (Libet et al 1979), which Klein argues may be too long for stimuli that are significantly greater than threshold. I do not wish to argue this point, since I am not aware of the relevant figures (the 100 msec figure referred to by Klein being not relevant to the situation I was considering, as far as I can see). A one-second's delay for a consciously controlled free-willed response seems already an inordinately long time.

13.7 I am not claiming that these considerations are decisive, in any way, as indications that the quantum/relativity puzzles concerning the nature of time and causality have significance for our consciousness and perception of "the flow of time". However, it seems to me that it is quite possible that there is something very odd going on concerning the timing of conscious events, if only for the reasons indicated in *Shadows* Section 7.11 that the role of time with respect to consciousness is quite different from its role in physics - in that it is only with the phenomenon of consciousness that time seems to "flow". I certainly hope that more experiments of the types that Libet and his associates have been performing will be carried out in the future. I suspect that there may be further surprises in store.

14. Some Remarks on Biology

14.1 Many people have expressed reservations (of widely differing degrees) concerning the biological speculations put forward in *Shadows*. I have referred (in Section 12 above) to Klein's worries about the difficulty of maintaining quantum coherence within individual microtubules and, moreover, of this quantum coherence straddling, simultaneously, a great number of microtubules within collections of separated neurons. I agree with Klein that it would be an extraordinary challenge to see how such organization might be achieved. Yet, I maintain that somehow Nature must indeed have accomplished this extremely remarkable task. In this section, I shall try to address this issue further, and also address some of the other objections that have come my way. I shall also relate several new things, in relation to these issues, that I have learned since I wrote *Shadows*.

14.2 One complaint that I have heard is that the biological purpose of microtubules within cells is "already known", namely that they are there to provide the "tracks" along which molecules (generally known as "organelles") are transported from one part of a cell to another; and they grow, shrink, or bend in ways that are designed to influence the movement of cells. Moreover, so these arguments might continue, their tubelike construction is to give them structural strength - so there is no need to ask for a separate purpose for the tubes, such as to isolate some kind of quantum-coherent activity taking place within the tubes, from the outside environment. I do not doubt that microtubules indeed perform the tasks that they are presently believed to perform, and many more besides. But that is no argument against their *also* serving the additional purposes that I require of them. We know of many instances where Nature uses the same structure for many different purposes. We know that mammalian noses, for example, filter substances from the air before it reaches the lungs (not to mention their importance to the sense of smell). Yet this is no argument against elephants *also* using their noses in delicate ways to pick up objects from the ground!

14.3 A more serious argument is the lack of direct evidence for the type of "cellular automaton" activity that Hameroff and his colleagues have been arguing for in patterns of tubulin conformations along microtubules. The existence of some kind of activity of this general nature is indeed part of the general picture that Hameroff and I would require for our model of the physical processes underlying consciousness. To obtain direct experimental support for this kind of activity would be a key issue, and I certainly hope that it will be possible to design experiments to test it. Experimental support for the existence of some kind of quantum coherence within microtubules is a matter of even greater importance for the ideas that I have been promoting in *Shadows*. There is no doubt that definitive experiments would be difficult to perform, especially since there is a distinct possibility that the relevant effects might require microtubules *in vivo* rather than *in vitro*. I have been informed by Guenther Albrecht-Buehler that there is some kind of coating (analogous to the myelin sheaths of neurons) that microtubules have *in vivo* which tends not to be present *in vitro*.

14.4 On the theoretical side, some progress has been made. Work by Tuszynski et al (1996) gives theoretical support for information processing (of the Hameroff type) to be possible, within an appropriate temperature range, provided that the microtubule possesses the structure of what is known as the "A-lattice", which is indeed the structure depicted in *Shadows* figures 7.4, 7.8, 7.9, on pp. 359, 363. However, the work of Mandelkow and Mandelkow (1994) indicates that many (perhaps most) microtubules seem to have a somewhat different structure, known as the "B-lattice", in which there is a "seam" running the length of the microtubule. Tuszynski et al argue that the B-lattice is not capable of sustaining Hameroff-type information processing, but it may well be appropriate for the transporting of organelles. It would be extremely interesting to have information about which kind of lattice structure is prevalent in axons, dendrites, non-neuronal cells, etc.

14.5 With regard to the theoretical possibility of quantum coherence within microtubules, the model of Jibu et al (1994) seems well-founded, in which super-radiance effects are anticipated within microtubules (analogous to the activity of a laser), where the electromagnetic field interacts with ordered water. For this process to occur, it would be necessary for the water within the tubes actually to adopt this ordered structure, and to be appropriately free of the wrong kind of impurities, such as chloride ions. (Apparently, sodium, calcium, magnesium, and potassium ions, in low enough concentrations should not disturb the ordering.) It should be mentioned, however, that the type of coherent activity that is anticipated in the model of Jibu et al may not be sufficient for my purposes. Though it is a necessarily quantum effect, it is not, as it stands, a *quantum-coherent* effect of the type that my arguments require. Genuine quantum coherence seems to be necessary in order that the quantum/classical borderline can be probed, where the (non-computable) effect of the missing OR theory can significantly make its mark. (This comment has relevance to Klein's query at the end of his Section 1. *Classical* coherence in the brain may well occur, but it does not provide an opening for non-computational activity, which I argue is a characteristic feature of consciousness.) The Jibu et al mechanism may be part (though not all) of what is needed.

14.6 An interesting possibility has come my way, which may conceivably have relevance to the question of how quantum coherence might get conveyed between one neuron and another (a question raised by Klein). As noted in *Shadows*, Figs. 7.11, 7.12 on pp. 365, 366, there are some particular molecules (clathrins) that inhabit synaptic boutons, which have the highly symmetrical structure of a *truncated icosahedron* (like a modern soccer ball). These clathrin molecules have importance in the release of neurotransmitter chemicals at synapses (whereby the nerve signals are transmitted from neuron to neuron). Although I do not have specific suggestions to make here, I am struck by the extraordinary symmetry of these molecules. It has been brought to my attention (by Roy Douglas, cf. Douglas and Rutherford 1995) that, according to the Jahn-Teller effect, such highly symmetrical molecules would have a large energy gap between the lowest quantum energy level and the next. This lowest level would be highly degenerate, and there would be interesting quantum-mechanical effects when this degeneracy is broken.

14.7 Energy gaps and symmetry breaking, of this general nature, are central to the understanding of superconductivity - and superconductivity is one of the few clear phenomena in which large-scale quantum coherence takes place. Known observationally since 1911, and explained quantum-mechanically in 1957, superconductivity had been thought originally to be an exclusively very low-temperature phenomenon, occurring only at a few degrees above absolute zero. It is now known to occur at much higher temperatures of -158 degrees Celsius, or perhaps even -23 degrees (although this is not properly explained). It does not seem to be out of the question that there might be similar effects at the somewhat higher temperatures of microtubules. Perhaps there are understandings to be obtained about the behaviour of microtubules from the experimental insights gained from such high-temperature superconductors.

14.8 Another question frequently asked is: what's so special about neuronal microtubules, as opposed to those, say, in liver cells? In other words, why isn't your liver conscious? In answer to this, it should be said that the organization of microtubules in neurons is quite different from that in other cells. In most cells, microtubules are organized radially, from a central region (close to the nucleus) called the *centrosome*. In neurons, this is not the case, and they lie essentially parallel with one another along the axons and dendrites. The total mass of microtubules within neurons seems to be much greater than in other cells, and they are mainly stable structures, rather than in most cells, where they continually polymerize and depolymerize (grow and shrink). Of course, there is much to be learned about the respective roles of microtubules in neurons and in other cells, but there does seem to be clear enough evidence for an essentially distinct role for (some of) those in neurons. (The A-lattice/B-lattice question would seem to be of importance here also.)

14.9 In this connection, I should mention something of considerable interest and relevance that I learned recently from Guenther Albrecht-Buehler (1981, 1991), which concerns the role of the *centriole*, that curious "T" structure (roughly illustrated in *Shadows*, Fig. 7.5, on p.360), consisting of two cylinders resembling rolled-up venetian blinds, constructed from microtubules and other connectingsubstances, which lies within the centrosome. In *Shadows*, I had adopted the common view that the centrosome acts in some way as the "control centre" of the cytoskeleton of an ordinary cell (not a neuron), and that it initiates cell division. However Albrecht-Buehler's idea about the role of the centriole is very different. He argues, convincingly, in my opinion, that the centriole is the *eye* of the cell, and that it is sensitive to infra-red light with very good directional capabilities. (Two angular coordinates are needed for identifying the direction of a source. Each of the two cylinders provides one angular coordinate.) Impressive videos of fibroblast cells provide a convincing demonstration of the ability of these cells to pinpoint the direction of an infra-red light source. This also provides some remarkable evidence for individual cells having considerable information-processing abilities, which is at variance with current dogma. One may well ask where the "brain" of a single cell might be located. Perhaps its structure of microtubules can serve such a purpose, but it does seem that the centrosome itself must have some central organizing role. In a single (non-neuronal) cell, the microtubules emanate from the centrosome. I gather from Albrecht-Buehler that the specific contents of the centrosome are not known. It seems that it would be important to know what indeed is going on in the centrosome. Does it have some information-processing capabilities? Is there conceivably some structure there that is capable of sustaining quantum coherence in any form? The answers to questions of this nature could have considerable importance.

14.10 I should make clear that I am not arguing for any consciousness (or consciousness of any significant degree) to be present for individual cells. But according to the views that I have been putting forward, some of the ingredients that are needed for actual consciousness ought already to be present at the cellular level. Individual cells can behave in strikingly sophisticated ways, and I find it very hard to see how their behaviour can be explained along entirely conventional (classical) lines.

14.11 All this notwithstanding, there is the question of whether microtubules are indeed necessary for consciousness to be present in human beings or other animals. An argument that I have heard presented - as though it were a conclusive refutation of this contention (cf. Grush and Churchland 1995, Edelman 1995) - is that the drug colchicine, which is given as a treatment for gout, depolymerizes microtubules, yet it does not influence the mental state; moreover, when colchicine is delivered directly to the brains of experimental animals, they appear to remain conscious. This can be answered (cf. Penrose and Hameroff 1995 and references contained therein) by pointing out that: (1) the blood-brain barrier is not significantly breached in the gout patients, so their neuronal microtubules are not disturbed, and (2) in any case, most brain microtubules, unlike those in non-neuronal cells, are stable structures that do not undergo cycles of polymerization and depolymerization, and are *resistant* to colchicine. A minority of neuronal microtubules - those involved in restructuring synaptic connections - are involved in such activity, however, and these could be affected by colchicine. Indeed, the experimental animals referred to above suffer a kind of "dementia", similar to Alzheimer's disease (Bensimon and Chernat 1991), a disease which has itself been linked to microtubule disruption (e.g. Matsuyama and Jarvik 1992).

14.12 Of course, there is the additional issue of how we could know whether a demented rat is or is not conscious. We must return to the question of what consciousness is and what are its external manifestations.

15. What is Consciousness?

15.1 In *Shadows* (cf. Section 1.12) I concentrated specifically on the quality of "understanding" as a particular manifestation of consciousness - a quality that would make its characteristic mark on external behaviour as well as being an internal manifestation of mentality. Only with respect to the quality of understanding have I been able to argue for non-computable ingredients being necessary. But, in my view, non-computable physical processes must also be essential for other aspects of conscious mentality.

15.2 Consciousness has its active aspects - basically the "free-will" issues that were considered in Section 13 above - and it has its passive aspects, which have to do with awareness and the vexed issue of *qualia*. Understanding fits somewhere between the two. In my view, anything that sheds some light on the problem of how a physical system can exhibit understanding must inevitably also shed some light on the "free-will" and "qualia" problems. Moreover, the issue of "understanding" seems to me to be one of the more tangible aspects of consciousness. I do not see how to say much that is scientifically useful about the qualities of "free will" or "awareness", but "understanding" is something that we can work with. Klein raises the issue of Wilczek's challenge of "looking for perceptual feats that humans can do more efficiently than robots". My answer would be: anything in which the quality of *understanding* is important. A good example is the chess problem presented in *Shadows* Fig. 1.7 on p.46. This problem was one of a series composed by William Hartston and David Norwood, consisting of chess problems, some of which were designed to be easy for humans but hard for computers (best solved using "understanding") and others, the other way about (best solved by "trying all possibilities"). This "Turing test" showed a virtually complete separation between humans and computers.

15.3 As must be clear from the preceding remarks, I do not believe that any real progress will be achieved towards solving the mysteries of how mental phenomena fit in with the physical universe until there are some important changes in our picture of physical reality. Perhaps these developments will lead to a theory in which "consciousness" finds some place within the purely physical descriptions of the world. One is reminded of such ideas as "panpsychism" (like those of Leibniz, Spinoza, or

Whitehead), where consciousness may play its part within the processes of physical action at its deepest levels. I do not have strong opinions as to the significance of such ideas, mainly because I have not studied them in detail. But I suspect the truth to have a much more compelling grandeur to it than any set of ideas that I have seen so far.

15.4 Thus, I certainly do not go along with the "no big deal" viewpoint of McDermott, a viewpoint which is, in effect shared by Baars when he takes the view that the solutions are all to be found within psycho-biology, without any significant change in our physical world-view being necessary. I think that Baars grossly underestimates the force of the arguments from logic. (And although, as a mathematician, I frequently make use of "variables", I have no idea what Baars means by "treating consciousness as a variable".) I fully accept that there are invaluable insights to be gained from the areas of psychology and biology. But, important though these areas are, I was not so much concerned with them as with physics in *Shadows*. For I believe that it is *also* fundamentally important to see whether our present physical world-view is in fact adequate for accommodating, in any way, the phenomenon of consciousness. I have extensively put the case here, and in *Shadows*, that it is not. The arguments from logic and from physics are not counter to those from psychology and biology, but complementary to them.

15.5 Likewise Moravec and McCarthy appear to belong to the "no big deal" school. McCarthy puts forward various suggestions for the circumstances under which he would consider that "consciousness" occurs. These are all within the computational model, so it is clear from this that I am not in agreement with him that his computer systems, acting according to his criteria, are *actually* conscious (in the sense that one could actually *be* such a system). Again, I fear that McCarthy does not appreciate the force of the logical arguments that I have given, which inform us that the quality of "understanding" cannot be accommodated within the computational model. It is easy to suggest *definitions* within the computational model (as McCarthy does) of such things as "consciousness", "awareness", "self-awareness", "intentions", "beliefs", "understanding", and "free will". But such definitions need not convey to us (and do not convey to me) any conviction that the corresponding mental qualities that humans actually possess are in any real sense captured by computational definitions of this nature. As I have argued extensively above, the actual quality of human *understanding* cannot be captured within *any* purely computational scheme. So it is clear that I cannot be in agreement with all of McCarthy's definitions.

15.6 Chalmers raises the issue of the distinction between "simulating", "evoking", and "explaining". I agree with him that there are indeed distinctions to be made. But I feel that these are distinctions that have more importance to a philosopher than to a scientist. Whilst I did myself distinguish between the possibility of "simulating" and of "evoking" consciousness in my "viewpoint B" and "viewpoint A" distinction in *Shadows* Section 1.3, I think that the normal scientific (as opposed to philosophical) stance would be to concentrate on what can be externally observed, so a system which succeeds in simulating the outward effects of consciousness would be suspected also of evoking them. Thus, my "viewpoint B" might not be a happy one for a hard-nosed scientist.

15.7 Likewise, a modern scientist might have trouble producing an "explanation" for something other than by producing a theory that in principle provides a (mathematical) "simulation" of that thing. It seems to me, therefore, that the "scientific" viewpoints (in this sense) are those who hold to the same position, with regard to computationalism, in respect of all three of the simulate/evoke/explain issues. And since Chalmers's "N" represents a denial of the competence of physics with regard to conscious mentality, this implies that we would be left with just "CCC" (computationalism all down the line) or "PPP" (non-computational physics all down the line). Thus, when I wear my scientist's hat, I am unable to understand why someone (such as Chalmers) can hold to different positions with regard to the simulate/evoke/explain issues, although when I wear my philosopher's hat, I can partly appreciate his point. However, I wear my scientist's hat much more frequently than my philosopher's hat!

15.8 But sometimes I try to wear both hats at once. The arguments in *Shadows* were concerned almost entirely with the "simulate" issue with regard to human mentality. I hope that those who study, with a genuinely open mind, the arguments given there (and the further discussions above) will come to accept that a non-computational physics will be needed in order even to simulate the actions of a conscious being. There are, in any case, powerful reasons for believing that profound changes in our physical world-view are in the offing. For the resulting science to be non-computable to the degree that seems to be required, we may well find the need for a science that is so different from the science of today that the *evoke* and *explain* issues with regard to mentality may finally find natural explanations.

Acknowledgement

I am grateful to the National Science Foundation for support under contract PHY 93-96246.

Notes

<1> As I understand it, there was to have been a tenth commentary, explicitly addressing a number of biological points, but unfortunately this article did not materialize. Nevertheless, in Section 14, I shall find it helpful to address a number of biological criticisms that have come to my attention.

<2> Most book-reviewers seem to have missed this argument too. In particular, Hilary Putnam, in his widely quoted review of *Shadows* in the Sunday New York Times Book Section (Putnam (1994) and reprinted, for some reason, in the Bulletin of the American Mathematical Society, Putnam (1995)) not only completely missed this argument, but tries to claim that I have not considered other issues that I have, in fact, discussed in great detail. The matters are thoroughly discussed again in Sections 6 and 7 of this reply.

<3> Dennett, in his 1995 book *Darwin's Dangerous Idea*, seems to be trying to make out that I do not believe that human abilities can have arisen by the process of natural selection, since I do not believe in computationalism. This is a very strange contention. Provided that the non-computational ingredients are present in nature, there is nothing against natural selection having made use of them - and it is my own contention that this is indeed how things were. If Dennett is arguing that I do not

believe that natural selection provides the sole explanation of the origin of human mentality, then he is right. In particular, some specific laws of physics are also needed, within whose scope natural selection must operate. But that, in itself, is hardly a radical position, nor an unscientific one!

<4> It appears that some people, on reading the section entitled "Contact with Plato's world" in Chapter 10 of *The Emperor's New Mind*, have picked up the curious view that I believe that mathematicians obtain their mathematical knowledge by use of some direct mystical quality not possessed by ordinary mortals (see Grush and Churchland 1995, for example), and even that I may be claiming for myself a particularly unique such quality! This is a complete misreading of what I had intended in that section; for I was simply trying to find some explanation of the fact that different mathematicians can communicate a mathematical truth from one to another even though their modes of thinking may be totally dissimilar. I was arguing merely that the mathematical truths that each mathematicians may be groping for are "external" to each of them - these truths being "inhabitants of Plato's timeless world". I was certainly not arguing for a fundamentally *particular* quality of "direct Platonic contact" to be possessed only by certain individuals. I was referring simply to the general qualities of "understanding" (or "insight") which are in principle available to all thinking individuals (though they may perhaps come somewhat more *easily* to come individuals than to others). These qualities are not mystical - but as Gödel's theorem shows, there is indeed something rather mysterious about them.

<5> See, for example, Grush and Churchland (1995), and for a reply, Penrose and Hameroff (1995).

<6> John Searle, in his interesting recent review of *Shadows* in the *New York Review of Books* (November 2, 1995), seems to be making a somewhat similar point. However, he does not appear to have grasped, properly, the key notion of *non-computability* (and the fact that it has observational manifestations). See, in particular, the discussions given here (Section 7, 10, 11, 15) and also Sections 3.2 onward of *Shadows*, which explicitly address the conscious/unconscious issue that Searle raises, with regard to an algorothmic basis for mathematical understanding. The aforementioned sections may serve to clarify my own position concerning what he claims are my "fallacies".

<7> F.J.Tipler, in a review of *The Emperor's New Mind*, used a similar "finiteness" argument, specifically referring to the (absurdly large) Bekenstein bound on the information that can be stored within an object of a given (say human) size. I explicitly addressed this argument from finiteness in my responses to Q7 and Q8 in Section 2.6 of *Shadows*. However, I had deliberately desisted from indicating the personal individuals whom I had in mind (in this case Tipler) as putting forward the various specific "queries" that I was responding to. There is a certain irony for me in this, because in his review of *Shadows* in *Physics World*, Tipler chastised me, claiming that I did "not even mention ... the Bekenstein bound" (Tipler 1994), not having even noticed this section of the book that had been specifically aimed at his own arguments!

References

Albrecht-Buehler, G. (1981). Does the geometric design of centrioles imply their function? *Cell Motility, 1,* 237-245.

Albrecht-Buehler, G. (1991). Surface extensions of 3T3 cells towards distant infrared light sources. *Journal of Cell Biology, 114,* 493-502.

Bensimon, G. and Chernat, R. (1991). Microtubule disruption and cognitive defects: effects of colchicine on learning behavior in rats. *Pharmacology and Biochemistry of Behavior, 38,* 141-5.

Davis, P.J. and Hersch, R. (1982). *The Mathematical Experience*. Harvester Press.

Deeke, L., Groetzinger, B., and Kornhuber. H.H. (1976). Voluntary finger movements in man: Cerebral potentials and theory. *Biology and Cybernetics, 23,* 99.

Dennett, D.C. (1995). *Darwin's Dangerous Idea.* New York: Simon and Schuster.

Devlin, K. (1988). *Mathematics: The New Golden Age*. London: Penguin Books.

Diosi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Physical Review, A40,* 1165-74.

Douglas, R.R. and Rutherford, A.R. (1995). Pseudorotations in molecules I: Electronic triplets. To appear.

Edelman, G. (1995). Quoted on p. 323 of *Frontiers of Complexity*, by Peter Coveney and Roger Highfield. London: Faber and Faber.

Feynman, R.P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.

Feynman, R.P., Morinogo, F.B., and Wagner, W.G. (1995). *Feynman Lectures on Gravitation*. Reading, MA: Addison-Wesley.

Gardner, M. (1970). Mathematical games: the fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American, 223,* 120-123.

Ghirardi, G.C., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review, D34,* 470.

Ghirardi, G.C., Grassi, R., & Rimini, A. (1990). Continuous-spontaneous-reduction model involving gravity. *Physical Review, A42,* 1057-64.

Grush, R. and Churchland, P.S. (1995). Gaps in Penrose's toilings. *Journal of Consciousness Studies, 2,* 10-29.

Hameroff, S.R. (1987). *Ultimate Computing: Biomolecular Consciousness and Nano-Technology*. Amsterdam: North Holland.

Hameroff, S.R. and Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules - a model for consciousness. In S Hameroff, A. Kaszniak and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.

Hameroff, S.R. and Watt, R.C. (1982). Information processing in microtubules. *Journal of Theoretical Biology, 98,* 549-61.

Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H., Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *BioSystems, 32,* 195-209.

Karolyhazy, F. (1966). *Nuovo Cim., A42,* 390.

Karolyhazy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat, 12,* 24.

Libet, B. (1992). The neural time-factor in perception, volition and free will. *Review de Metaphysique et de Morale, 2,* 255-72.

Libet, B., Wright, E.W. Jr., Feinstein, B. and Pearl, D.K. (1979). Subjective referral of the timing for a conscious sensory experience. *Brain, 102,* 193-224.

Mandelkow, E-M. and Mandelkow, E. (1994) Microtubule structure. *Current Opinions Structural Biology, 4,* 171-179.

Matsuyama, S.S. and Jarvik. L.F. (1992). Hypothesis: Microtubules, a key to Alzheimer's disease. *Proceedings of the National Academy of Science USA, 86,* 8152-6.

McDermott, D. (1990). Computation and consciousness. *Behavioral and Brain Sciences, 13(4),* 676.

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind.* Oxford: Oxford University Press.

Penrose, R. (in press). On gravity's role in quantum state reduction. *General Relativity and Gravitation.*

Penrose, R. and Hameroff, S. (1995). What 'gaps'? - reply to Grush and Churchland. *Journal of Consciousness Studies, 2,* 99-112.

Poundstone, W. (1985). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge.* Oxford: Oxford University Press.

Putnam, H. (1994). The best of all possible brains? *New York Times Book Review, Nov. 20,* 7-8.

Putnam, H. (1995). Review of *Shadows of the Mind*, by Roger Penrose. *Bulletin of the American Mathematical Society, 32,* 370-373.

Tipler, F.J. (1994) Can a computer think? Part II. *Physics World, December,* 51-52

Tusznyski, J., Trpisova, B., Sept, D., and Sataric, M.V. (1996). Microtubular self-organization and information processing capabilities. In S. Hameroff, A. Kaszniak, and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.

245

CHANDRASEKHAR, BLACK HOLES, AND SINGULARITIES

J. Astrophys. Astr. **17** (1996), 213–31

This article was written in honour of Subrahmanyan Chandrasekhar, who died in 1995. There are a few awkward typographical points. On p. [219], 7 lines up, the equation is spaced badly; on p. [220] there are two occurrences of the symbols *I*– and *I*+, which should be $\mathscr{I}^-$ and $\mathscr{I}^+$, respectively.

J. Astrophys. Astr. (1996) **17**, 213–231

Chandrasekhar, Black Holes, and Singularities

Roger Penrose
Mathematical Institute, University of Oxford, Oxford 0X1 3LB, UK.

1. Historical background

The dilemma that was presented to the scientific world by Chandrasekhar's early work (1931) on the existence of a maximum mass for white dwarf stars took some while to be fully appreciated. There were some, such as Eddington, who did seem to understand the alarming implications of Chandra's conclusions. Assuming the correctness of the relativistic equations of state, it seemed that a white dwarf star of mass more than about 1.4 of a solar mass would have to collapse inwards, its density increasing indefinitely as the body approached a singular configuration at the centre. However, Eddington himself regarded this as a *reductio ad absurdum*, concluding, instead that there must be something wrong with Chandra's use of the relativistic equations. Eddington supposedly had in mind that some new physical principles must come into play in order to save the star, perhaps such as embodied in his own approach to a deeper fundamental theory (Eddington 1946). Taken at the level of phenomena at which Chandra's discussion was intended to apply, there is no doubt that Chandra's analysis was the correct one, as the experts seemed to have appreciated even at that time (at least privately), despite the weight that Eddington's authority attached to the contrary view. Yet Eddington had a point: the impossibility of an equilibrium state would lead to the star's unstoppable collapse. Would this collapse continue until the star becomes so compressed that it reaches its Schwarzschild radius ($r = 2m$), thought at that time to be a dimension at which the very metric structure of space-time becomes singular? In any case, as the star continues to collapse radially inwards, it appears that it should reach a state where the density becomes infinite. And, according to Einstein's general theory of relativity, infinite matter density would in itself imply a singularity in space-time structure. Thus, it would seem that even if one accepted the procedures of conventional physics completely, with the conclusions that Chandra so strongly argued for, the unending collapse of the star would lead to a situation in which those very procedures would ultimately have to be abandoned.

By nature conservative in his approach, Chandra would certainly not have been attracted to a radical kind of speculation that had so occupied Eddington in his later years. In the 1930s, even general relativity was not regarded as a worthy activity for aspiring astrophysicists to devote themselves to. In any event, for whatever reason, Chandra chose not to mount a direct attack on the problem of gravitational collapse.

In that regard, he bided his time. Indeed, there were many other issues that would occupy his attentions for a quarter of a century!

In around 1960, having completed his works on stellar structure and stellar dynamics, on radiative transfer, and on the equilibrium or stability of various structures of astrophysical interest, he finally resolved to enter into a study of general relativity. He attended the 1962 International Conference on General Relativity and Gravitation in Warsaw as a 'student', in order to attain an overview of current research in the subject. Even after he had thoroughly prepared himself, he did not directly address the issue of the fate of a collapsing star. He was more concerned, first, with the effects of general relativity on the stability of gravitating bodies, and then with general-relativistic corrections to the Newtonian dynamics of collections of masses, including the effects of gravitational radiation damping. These studies were relevant to the issue of how considerations of general relativity might affect the onset of gravitational collapse, not with the result of the collapse itself. In short, the overall conclusion was that the influence of general relativity accelerates gravitational collapse somewhat, over and above the various influences of other physical effects.

So was collapse right down to a singularity of some sort the inevitable conclusion? For a white dwarf star, there might still be a number of other possibilities open to it. It might, for example, have the opportunity to find ultimate rest as a neutron star. But it was already an implication of Chandra's analysis (cf. Landau's simplified approach of 1932) that it applied also to the relativistic equations of state for neutron matter. The analysis was carried out in detail by Oppenheimer and Volkoff (1939), the conclusion at that time being that the maximum mass for a neutron star was a little smaller, even, than for a white dwarf. Allowing for the fact that the neutrons themselves can become converted to other massive particles (not known at the time), somewhat larger limiting masses were subsequently suggested. On the other hand, very general considerations, based on fundamental principles such as causality, allowed the conclusion to be made that there was an absolute overall limit of not much more than 3 solar masses. There appears to be no clear agreement as to what the actual limit is, but observations seem to suggest something like about 1.4 solar masses. Accordingly, no final solution to the problem of gravitational collapse is to be found along these lines.

Already in 1939, Oppenheimer and Snyder had faced this problem squarely, considering the situation of a collapsing spherically symmetrical dust cloud, of uniform density, treated according to Einstein's general relativity. In their description, they provided the first explicit model of a collapse to what is now called a 'black hole'. They showed that although the so-called 'Schwarzschild singularity', which occurred at radius $r = 2m$, is not actually a singularity –what is now called the 'horizon' – there is still a space-time singularity at the centre, where the density of the dust indeed becomes infinite. Ironically, it was Eddington himself who first published, in 1924, a metric form for the spherically symmetric Schwarzschild space-time in which the 'Schwarzschild singularity' is exhibited in its true nature as a null hypersurface (actually two null hypersurfaces). However, Eddington appeared not to appreciate what he had himself done. Before Oppenheimer and Snyder, Lemaȋtre (1933) had appreciated that the 'Schwarzschild singularity' could be locally eliminated by a coordinate change but he had not provided an overall picture of collapse to a black hole. Later Synge (1950) and others found the complete extension of the Schwarzschild solution.

Since there is still a singularity in the Oppenheimer-Snyder collapse model (at $r = 0$), the Chandrasekhar dilemma is not removed by their collapse picture. However many people remained unconvinced that this description would necessarily be the inevitable result of the collapse of a star too massive to be sustainable as either a white dwarf or neutron star. There were a number of good reasons for some scepticism. In the first place, the equations of state inside the matter were assumed to be those appropriate for pressureless dust, which is certainly far from realistic for the late stages of stellar collapse. Moreover, the density was assumed to be constant throughout the body. With realistic material, there are many alternative evolutions to that described by Oppenheimer and Snyder. For example, nuclear reactions set off at the centre could lead to an explosion – a *supernova* - which might perhaps drive off sufficient mass from the star that a stable equilibrium configuration becomes possible.

Most serious of all was the assumption of exact spherical symmetry. Since, in this picture, all the material of the body is aimed directly at the central point, a resulting density singularity could easily be the result merely of this fortuitous focussing. It could be expected that the introduction of even the slightest perturbation away from spherical symmetry might cause most of the inward falling particles of the body's material to miss the central point, so that even though the density might get very large there, it might well not diverge to infinity.

Indeed, as late as 1963, the Russian school of Lifshitz and Khalatnikov (1963) had claimed that the 'generic' solution of the Einstein equations ought to be free of singularities. However, it became evident that there must be a significant error in their work when the early singularity theorems were established (Penrose 1965; Hawking 1966; Hawking & Penrose 1970). In fact, such an error was found and corrected by Belinskii (cf. Belinskii, Khalatnikov and Lifshitz 1970, 1972), their final conclusion being that singularities could occur in generic solutions after all. (Lifshitz and Khalatnikov had omitted a certain degree of freedom in their expansions.)

In any case, it is difficult to form any firm conclusions about the final stages of gravitational collapse from the kind of power series analysis that the Russian school had adopted. Perturbation analysis is not well suited to the extreme situations that would be expected to arise when a space-time approaches a singular state. Moreover, the possibility of finding explicit solutions of sufficient generality to describe what happens in a realistic collapse could be virtually ruled out. For reasons such as these, the approach adopted in the singularity theorems was completely different. Instead of attempting to work out what happens in detail to a collapsing star, general overall considerations, of a largely topological nature were employed so as to derive general properties of the solution. In essence, the conclusions were all of a negative character, in the sense that they ruled out certain things, showing that they cannot happen – rather than showing what actually does happen.

What cannot happen, according to these theorems, is an ordinary singularity-free evolution – assuming that equations of state satisfy a reasonable energy – positivity condition – if the collapse reaches a certain point of no return. This 'point of no return' can be characterized as the existence, in the space-time, of what is called a 'trapped surface'. This is a compact spacelike 2-surface whose null normals converge into the future (which means that if a flash of light were to be emitted at the surface, then the area of any element of cross-section of the emitted light rays must decrease; cf. Penrose 1965). Another slightly different way of specifying an appropriate 'point of no return' is the existence of a point in the space-time through which all light rays into the future begin to reconverge somewhere (cf. Hawking & Penrose 1970). By an 'ordinary singularity-free evolution', I mean that it must be possible to continue the space-time (non-singularly) into one in which all null and/or timelike geodesics have infinite affine length and for which appropriate causality conditions are maintained (such as the existence of a global Cauchy hypersurface or merely the absence of closed timelike curves together with a condition that the space-time is in some mild sense 'generic'). There are slight differences in the details of these various conditions, depending upon which singularity theorem is being appealed to. (For further information, see Hawking & Penrose 1970; Hawking & Ellis 1973).

The upshot of all this is that once a trapped surface (or reconverging light cone) has formed, there is no way, within the scope of existing physical laws, to extend the space-time indefinitely. This is the ultimate dilemma that Eddington was, in effect, shying away from. But why should we expect a trapped surface to arise in any case? In the Oppenheimer-Snyder picture, just after the collapsing matter shrinks within the Schwarzschild radius at $r = 2m$ (the event horizon), there are trapped surfaces in the space-time region immediately surrounding the body. Since the trapped-surface condition is, by its very nature, something undisturbed by (adequately small) finite perturbations, the essential issue is whether or not a collapsing body, or collection of bodies is ever likely to reach the vicinity of its Schwarzschild radius – and just beyond. Chandra's work on the influences of general relativity on the stability of massive bodies, let alone his much earlier initiation of the entire line of thinking concerning the maximum mass of bodies held apart by relativistic degeneracy pressure, provided strong support for the view that unstoppable collapse to the neighbourhood of the Schwarzschild radius was probable.

Nevertheless, these arguments are not totally convincing, owing to the lack of complete information about details of the internal state of a star, where the density might exceed, by several orders of magnitude, even that inside an atomic nucleus. However, if it is matters of principle that we are concerned with, these details are not important. One can envisage situations in which trapped surfaces will arise even for densities that are as low as, say, that of air. By making the total mass of the collapsing system large enough, the density at which the body crosses its

Schwarzschild radius can be made as small as we please. At the centres of large galaxies there would be collections of ordinary stars which, if they were to find themselves in a small enough region all at once, would be surrounded by a trapped surface even though the stars are not yet in contact, so the total density is less than stellar density. This follows from very basic considerations of general relativity – even directly observed ones – concerning the focussing power of mass density on light rays. Rather easier than the trapped- surface condition is to use the condition of a reconverging light cone. Suppose that the radius of the region is 10^4 km. The light rays emerging from a point at the centre of the star cluster would encounter sufficient stellar material that they indeed begin to reconverge, as follows from a simple order-of-magnitude calculation. (See Penrose 1969, p.266, for a simple description of this idea.)

It follows from this, and the singularity theorems, that if conventional physical ideas hold true, we are forced into having to face up to the occurrence of space-time singularities. We must ask what is the nature of these singularities and what are we to do about them. In fact, the singularity theorems are almost completely silent about the nature of the singularities themselves. The theorems are simply existence theorems, and say almost nothing about the location of the singular regions, let alone anything about their detailed nature. From later work (e.g., Clarke 1993, and also the earlier work of Belinskii, Khalatnikov and Lifshitz 1970, 1972) there were certain very incomplete indications as to the singular behaviour of the Weyl curvature tensor. Enough is known to suggest that the structure of the singularities in a generic collapse will be quite different from that which is encountered in the big bang (divergent Weyl curvature in the former, essentially vanishing Weyl curvature in the latter – a feature intimately connected with the second law of thermodynamics, cf. Penrose 1979). But very little is known in detail. As to 'what we are to do about them', it is clear that these singularities take us outside the domain of classical general relativity. Without the appropriate union between general relativity and quantum mechanics being to hand, we are presented with an impasse.

The existence of singularities does not, however, imply the existence of black holes. This deduction requires the additional assumption of what is called 'cosmic censorship'. Cosmic censorship (Penrose 1969, 1978) asserts that naked singularities will not occur in a generic gravitational collapse. Such a singularity, roughly speaking, would be one that could be seen by an outside observer. Cosmic censorship would imply that the region lying to the future of all singularities resulting from gravitational collapse cannot reach future null infinity $I+$. The boundary of the entire region which cannot be connected to $I+$ by a causal (i.e. timelike or null) curve defines the (absolute) *event horizon H*. We thus see that a gravitational collapse resulting in singularities subject to cosmic censorship will lead to the existence of a *horizon H* which 'shields' all the singularities from directly revealing themselves from the outside world. This horizon describes a black hole. (See Penrose 1973)

Various theorems (Israel 1967; Carter 1970; Robinson 1975; Hawking 1972) can now be called into play, where it is assumed that the Einstein vacuum (or Einstein-Maxwell) equations hold from the neighbourhood of H out to infinity, after the collapsing body has fallen through. The upshot of these theorems is that if the space-time in this region settles down to a stationary configuration, then it is described by the Kerr metric (Kerr 1963), completely defined by two parameters m and a (or, in the Einstein-Maxwell case, the Kerr-Newman metric (Newman *et al.* 1965), defined by three parameters m, a, and e).

2. Chandra's work on perturbations of black holes

It is a remarkable fact that black holes, when they settle down to an exactly stationary configuration have such a precise and explicit mathematical description. Chandra never ceased to be impressed by this fact, remarking at the beginning of his Prologue to *The Mathematical Theory of Black Holes* (1983):

> "The black holes of nature are the most perfect macroscopic objects there are in the universe: the only elements in their construction are our concepts of space and time. And since the general theory of relativity provides only a single unique family of solutions for their descriptions, they are the simplest objects as well."

Chandra's study of black holes was based upon these stationary cases: the Kerr (and Kerr-Newman) space-times – together with the Schwarzschild (and Reissner-Nordström) specializations. Since these stationary metrics were known explicitly, he was able to study, in comprehensive detail, the first-order gravitational perturbations away from the stationary Kerr configuration – and gravitational-electromagnetic ones in the Reissner-Nordström case. Moreover, he provided a comprehensive treatment of the Maxwell and Dirac equation on a Kerr background. In all these cases, he found remarkable algebraic/differential relations which enabled him to separate and decouple the equations.

Chandra was a relative latecomer to the study of black holes. In his early work on white dwarf stars and the inevitability of their collapse when too massive, leading to his early realization of the dilemma referred to in section 1, Chandra had been far ahead of his time. But his assault upon the very problem that his early researches had thrown up was delayed until after most of the groundwork had been carried through by others. There were, of course excellent reasons for this. Chandra's monumental work on other topics had to be completed first. It was his way of working that he would devote himself in a single-minded way to one topic at a time, thereby achieving· the phenomenal thoroughness and depth that he strove for in each topic.

However, he also had a long-standing interest in general relativity and, specifically, in what came to be known as black holes. But until about 1974 (about the time when most experts in the subject had moved away from the classical theory and were turning to the implications of Hawking's discovery of black-hole radiation), he had not felt ready to embark upon his detailed assault upon the area of black holes.

In particular, he felt that he needed to complete what had been his most recent work – on the stability of rotating stars – before doing so. As it happens, in this work, he was studying perturbative general-relativistic effects, including the effects of gravitational radiation. Thus, this research provided him with a natural route into the study of perturbations of black holes. Indeed, the various researches that Chandra has undertaken should not be thought of as being totally independent of one another. In particular, in his rotating star stability work, as in so many other things that he achieved, he exploited his extraordinary ability with equations, and he already had much of the basic framework of ideas to hand, ready for his concerted attack on the problems of black-hole perturbation. Indeed, it appears to have been Teukolsky's separation of the equations for gravitational perturbations of a Kerr black hole (Teukolsky 1973) that stimulated Chandra's actual entry into the subject.

The physical motivations for the study of black hole perturbations came from a desire to know how such an object, initially in a stationary configuration, would react if slightly disturbed. Its response could involve the emission of gravitational waves, with another part of the disturbance disappearing into the hole. It would be argued that this response would be governed primarily by the structure of the black hole itself. The linear perturbations of the hole would be described by linear equations, and could therefore be analysed in terms of the appropriate 'modes'. These are not quite like the modes of vibration of a perfectly elastic body, because of the effective dissipation that occurs both because of gravitational radiation and through. loss into the hole itself. The frequencies are therefore complex, with imaginary parts describing the decay of the modes.

Chandra's first paper on black-hole perturbations (Chandrasekhar 1975) was concerned with the Schwarzschild black hole. At that time there already existed a thorough analysis of these perturbations, dating back to the work of Regge and Wheeler in 1957. In particular, Zerilli (1970) had shown that, splitting the perturbation into different components for the individual spherical harmonics, each component satisfies a Schrödinger equation, corresponding to that for a wave of time dependence $e^{i\sigma t}$ (with complex frequency σ having an imaginary part describeing the damping rate) and spatial dependence $Z(x)$, impinging on a potential barrier defined by a certain smooth potential function $V(x)$:

$$\left(\frac{d^2}{dx^2}+\sigma^2\right) Z = VZ.$$

The variable x is related to the standard Schwarzschild radial coordinate r by $x = r + 2m \log(r - 2m)$. The potential function V is a particular explicit function of r , depending on the mass m and the choice of spherical harmonic. (Units are always chosen so that the speed of light c and Newton's gravitational constant G are both unity.) Accordingly, the general perturbation problem for the Schwarzschild black hole can be reduced to that of finding the transmission and reflection coefficients of a simple one-dimensional barrier-penetration problem in quantum mechanics. On the other hand, Bardeen and Press (1973) had obtained a different set of equations to describe the same perturbations, and various relations between all these expressions and procedures had remained somewhat mysterious. Rederiving all these expressions in his own way, Chandra was able to understand and explain several of these mysterious features, most notably a curious relationship between the coefficients that arise from the odd- and even-parity (or, as Chandra preferred, axial and polar) perturbations.

Later, he sought to analyse the deeper reasons underlying such relationships, noticing that one could understand these in terms of the procedures of inverse scattering, according to which consistency conditions of the nature of the Korteweg-de Vries equation arise (Chandrasekhar 1982). In this way, Chandra provided an opening into the intriguing and mathematically fruitful area of integrable systems, and it is likely that the last word on these matters has by no means been heard (see comments in section 4).

The initial disturbance which causes a perturbation to the geometry of a stationary black hole could be caused by some physical object falling into the hole, or it could be taken to have the form of incoming gravitational (or gravitational-electromagnetic) waves, impinging on the hole from outside. Particularly in his later work on the subject, Chandra preferred to emphasize the latter viewpoint, this having the advantage that no foreign ingredients are imported into the system of equations under consideration, everything being described in terms of the Einstein vacuum equations (or Einstein-Maxwell equations). Of course, in an actual astrophysical situation, there would normally be some other source for a significant black-hole perturbation, but from the point of view of making the mathematical treatment self-contained, this approach has considerable advantages.

In a situation of this nature, there is an incoming component (from *I*–) and two 'outgoing' components, one which escapes out to infinity (*I*+) and another one which falls into the hole. As mentioned above, the situation is closely analogous to that which arises with one-dimensional potential scattering and barrier penetration in ordinary quantum mechanics – and for each separate spherical harmonic, the mathematical description is precisely of this form. There is an incoming wave train and an outgoing reflected wave train, accompanied by the part which passes through the potential barrier. In particular, there will be that situation which arises when the incoming influence is 'switched off' and the black hole 'rings' according to its natural frequencies. These modes are what are called 'quasi-normal modes', characterized by the fact that there is no component coming in from *I*– (and no component coming out from the interior of the hole – a geometrical impossibility in any case unless one allows for a delayed burst of radiation from the collapsing matter which originally produced the hole, which would have to have hovered, exponentially decaying, at the hole's horizon since its formation). Thus, each quasi-normal mode is composed only of a wave train escaping to *I*+ and a wave train falling into the hole. Each of these would be exponentially decaying modes (but badly behaved at spatial infinity owing to their exponential blow-up at infinite negative times). Chandra studied these modes for the Schwarzschild black hole in his 1975a paper with Detweiler (and, later, for the Kerr black hole), but he warned against too much reliance in these giving a complete characterization of the decay behaviour of black-hole perturbations and on too much faith being placed on the analogy with normal modes of an elastic material body. As far as I am aware, there are still unanswered questions concerning completeness and other issues for quasi-normal modes for black holes.

After his comprehensive treatment of the spherically symmetrical (i.e., Schwarzschild) case, Chandra moved on to his study of perturbations of the Kerr space-times. These possess axial symmetry and include angular momentum. It was, after all, the remarkable simplicity of general relativity's implication that stationary vacuum black holes have to be Kerr (or Schwarzschild) metrics that so attracted him to this area of research. The fact that Teukolsky (1973) was able to separate the radial parameter and the angular coordinate θ to obtain a pair of decoupled equations for the gravitational perturbations of the Kerr metric was a surprising additional bonus. Chandra rederived the Kerr metric in his own way (Chandrasekhar 1978a). With Detweiler (Chandrasekhar & Detweiler 1975b), he showed that Teukolsky's equations for the gravitational perturbations can be reduced to a (complex) one-dimensional wave equation of the type considered in the displayed equation above, with four possible potentials (but now depending on the frequency σ); moreover, they showed that the reflection and transmission coefficients are the same in each of the four cases, thereby illuminating some puzzling relationships. This work was continued in an important series of papers (Chandrasekhar 1976a,b, 1978a,b,c, 1979a,b, 1980, 1983) in which Chandra completed a very thorough and detailed analysis of the gravitational perturbations of a Kerr black hole. These papers are full of remarkable algebraic and differential relationships and identities, following on from those that had been established by Teukolsky and Starobinski.

Teukolsky (1973) had also separated the equations for scalar waves and electromagnetic fields on a Kerr background. Chandra treated both of these fields in two papers in 1976, again reducing the equations to the same type of one-dimensional wave-equation as above. He then showed (Chandrasekhar 1976c), ingeniously performing separation prior to decoupling, that the Dirac equation for the electron could also be separated and decoupled in a Kerr background. Teukolsky's work had already covered the case of a neutrino field

$$\nabla^{AB'} \nu_{B'} = 0,$$

but for a Dirac field with non-zero mass $2^{1/2}$ μ in 2-component spinor form, we have the coupled pair

$$\nabla_{AA'} P^A + i\mu\, \bar{Q}_{A'} = 0$$
$$\nabla_{AA'} Q^A + i\mu\, \bar{P}_{A'} = 0$$

and the separation of these had proved to be a stumbling block. Chandra took special delight in the fact, as noted in half of a sentence at the end of his paper, that by taking the limit in which the black hole's mass is set to zero, one obtains the separation of Dirac's equation in oblate spherical coordinates in flat space-time – a feat which had not been achieved before.

In his study of the Dirac equation, Chandra showed that he had mastered the 2-spinor formalism, which is not particularly familiar to physicists generally, there being an almost universal tendency for them to phrase their discussions of the Dirac equation in terms of the (superficially simpler but ultimately more complicated) 4-spinor formalism. In this, and also in much of his work in gravitational and electromagnetic perturbations (and in his later work on colliding plane waves), Chandra exhibited a great facility with what has become known as the 'NP-formalism' (the method of spin coefficients), which with Ted Newman and I had developed in 1962 to handle general relativity - effectively by combining the 2-spinor calculus with a Ricci-rotation-coefficient type of formalism (Newman & Penrose 1962). Chandra had specifically invited me to give a series of lectures in Chicago, in the early 1970s, primarily on the subject of this formalism, but I had in no way anticipated the powerful use that he would ultimately make of it.

To this, I might add a personal note of some irony. Some time later, in response to a question from me as to why he had not gone further and adopted the somewhat more streamlined later GHP formalism (Geroch, Held and Penrose 1973) Chandra had remarked that this would not simplify his equations in the way that I had imagined, because of an awkward problem about normalizing the spinor dyads against one another. Consequently, when writing the book *Spinors and Space-Time* with Wolfgang Rindler, I went to some trouble developing a generalized version of the GHP formalism in which the normalization condition was removed. This resulted in some extra complications which I know caused certain of my colleagues some irritation. However, I was not deterred by this, partly because the original GHP formalism can be extracted from this without difficulty, but more particularly because I believed that this was all in an excellent cause because Chandra could then directly incorporate this extended formalism into simplifying his equations! Apparently Chandra did not agree that this was any help, so we are left with a formalism still (as far as I know) looking for a good application. (Actually, I am still not convinced that it cannot be used with effect in the kind of thing that Chandra was doing. Perhaps some brave soul will have a look at it sometime.)

Chandra's work in relation to the Kerr metric applied to a background space-time in which the Einstein vacuum equations hold. Returning to the case of spherical symmetry, but where now the presence of an electromagnetic field is allowed for, Chandra studied gravitational-electromagnetic perturbations away from a stationary Reissner-Nordström black hole (Chandrasekhar 1979b; Chandrasekhar & Xanthopoulos 1979). Again, separation and decoupling of the perturbations occurs, with apparently rather little - but sometimes subtle – change required from the vacuum Schwarzschild case.

There is, however, one important change which does take place when one passes from the Schwarzschild to the Reissner-Nordström black hole, a change which occurs also when one passes to the Kerr black hole, this being the acquisition of a Cauchy horizon. A space-traveller who falls into the hole would, upon crossing the inner (Cauchy) horizon, enter a 'new universe' were it not for the fact that a shell of infinitely intense radiation coming from outside the hole would be expected to be encountered there. Some early work (cf. Simpson and Penrose 1973; McNamara 1978a,b) had provided a strong indication of this, but a more complete analysis of the perturbation theory, provided by Chandra with James Hartle in 1982, convincingly demonstrated the inevitability of this phenomenon (Chandrasekhar & Hartle 1982).

The combination of angular momentum with electric charge in a black hole gives rise to the solution of the Einstein-Maxwell equations known as the Kerr-Newman Metric (Newman *et al.* 1965). Chandra also studied Kerr-Newman black holes and their perturbations, but he was disappointed that he was unable to separate the equations for the gravitational perturbations, and he finally set the problem aside. If Chandra was not able to do it, that in itself would seem to be reason enough to believe that separation is not actually possible. However, the analogies with the Kerr and Reissner-Nordström black holes are very strong, so it is hard to resist the temptation that some separation, perhaps of a much more complicated nature, might lie somewhere behind the scenes. (In connection with this, following Chandra's separation of the Dirac equation on a Kerr background, Page was able, in 1976, to extend this result to the Kerr-Newman background.) I shall return briefly to the question of separation in section 4.

As was his general method of working, Chandra virtually set the seal on nearly ten years of research into this topic – black holes and their perturbations – by writing a superb book on the subject: *The Mathematical Theory of Black Holes*, published by the Clarendon Press, Oxford in 1983. Unlike the situation with his earlier books, however, he did not leave black holes entirely alone after that (cf. Chandrasekhar 1984, 1989; Chandrasekhar & Xanthopoulos 1989; Chandrasekhar 1990; Chandrasekhar & Ferrari 1990). Moreover, he certainly did not leave the subject of general relativity aside, as we shall see.

3. Colliding plane waves

It appears that the initial impetus that led to the next stage of Chandra's work, namely that on colliding plane waves in general relativity, was a letter from Yavuz Nutku (who had many years earlier been one of his students) which pointed out that the metric which arises when two impulsive gravitational waves with non-parallel polarization collide is described by the simplest solution of equations that Chandra had encountered in his derivation of the Kerr metric (Chandrasekhar 1978a).

There seems to be little doubt that it was Chandra's fascination with the mathematics of black holes that gradually began to turn him from a rigorous requirement that his work be directly relevant to astrophysically realistic situations. Of course, his study of black-hole perturbations was indeed astrophysically relevant, particularly because gravitational wave detectors may, before too long, possibly be able to detect the 'ringing' of a black hole after its formation or after it swallows a companion body. However, as his work continued, he became more and more seduced by the remarkable *mathematical* quality of the equations that he encountered.

Perhaps his 1989 paper with Xanthopoulos on two black holes attached to strings is the most extreme – and least 'Chandra–like' of his publications, in its departure from astrophysical realism. But colliding impulsive plane waves are also somewhat unrealistic. Since a plane wave must extend all the way to infinity, one cannot expect such a wave to be realized accurately in the physical universe. Moreover, impulsive gravitational waves (where the curvature tensor has the form of a Dirac delta function) do not provide a very reasonable idealization to the distant gravitational field of a violent event (such as, say, the congealing of two black holes) owing to the fact that there is an unending constant flux of gravitational radiation energy after the impulsive wave has passed. In addition, the nature of the space-time singularity that arises after the encounter between two such impulsive waves may have a special and unrepresentative structure, owing to its arising from a situation in which there is exact symmetry and precise focussing.

All this notwithstanding, colliding waves may well have provided an ideal framework for Chandra to make substantial progress towards an understanding of the very problem that his early researches into the equilibrium of white dwarfs had led to: the existence of *spacetime singularities.* Colliding plane waves indeed have the habit of leading to such singularities; and there is at least the possibility that these singularities may be closer to being realistic than those which occur in the Schwarzschild black hole (too special symmetry) and the Kerr black hole (closed timelike curves and intervening Cauchy horizon). It is not clear to me that Chandra was much concerned by the issue of physical realism at this stage in any case. Here was a family of solutions that he could study in detail. He could bring all of his magical gifts with equations to bear on these examples and, with luck, some deeper understanding of the nature and formation of space-time singularities could indeed come about.

His first paper on this subject described work that he and Valeria Ferrari had done (Chandrasekhar & Ferrari 1984) concerning the very situation that Nutku had described in his letter (Nutku & Halil 1977) [and which generalized the situation that Khan and I published in 1971 where the planes of polarization of the approaching impulsive plane waves were taken to be parallel (Khan & Penrose 1971)]. Their paper provided a very comprehensive analysis of this (Nutku-Halil) space time, including detailed expressions for all the NP spin-coefficients and curvature quantities.

In his next three papers on the subject, this time joint work with Xanthopoulos (Chandrasekhar & Xanthopoulos 1985a,b, 1986a) he included various matter terms into the equations: electromagnetic, perfect fluid (with $\varepsilon = p$), and null dust. Again the treatment was very detailed, but there were certain complicating issues that arose. For example, it was not reasonable to admit a delta function in the electromagnetic component to the incoming plane waves. For a delta function in the electromagnetic field would lead to a square of a delta function in the energy-momentum tensor and therefore in the Ricci curvature, which is not allowable. For there to be a delta function in the Ricci tensor, to accompany the already present delta function in the Weyl tensor describing the gravitational impulse, the Maxwell tensor would have to have the form of a square-root of a delta function. Chandra circumvented this particular problem by having the Maxwell field have only a step function at the gravitational impulse.

One source of potential confusion arose from the unusual way in which Chandra tended to obtain his solutions to the various field equations for this situation. The figure (taken from Chandrasekhar & Xanthopoulos 1986a) represents a space-time diagram for those two coordinates for which there is dynamic evolution. These are taken to be two null coordinates u and v (whose sum and difference may be regarded as the 'time' and the 'distance in the direction of propagation'), while the remaining two coordinates describe the flat plane surfaces of the waves – these planes being orbits of the two commuting translational symmetries of the space-time. Region IV is the flat spacetime between the approaching impulsive waves. Regions II and III are the portions of the two waves to the far side of the leading impulse. Region I is where the scattering between the waves takes place, this being the only region where serious work needs to be carried out in solving equations.

The flat portion of the space-time, prior to the arrival of either wave, is region IV. Chandra's ease with equations made it natural for him to start with the scattering region, and then to see what kind of waves in regions II and III might give rise to whatever solution for region I he might have found! This may seem strange to those physicists who think in terms of the evolution from Cauchy data on an initial hypersurface. In Chandra's case of the perfect fluid with $\varepsilon = p$ (the case illustrated in the figure), it turned out that in order to obtain the required solution for region I, the material in regions II and III could be taken to be null dust. The apparent difference between the equations of state holding in region I and those holding in regions II and III caused some confusion and a certain amount of controversy. Basically this was, I believe, a misunderstanding of Chandra's position concerning this situation. In fact, the perfect fluid with $\varepsilon = p$ has null dust as a limiting configuration, so there is not really any inconsistency between the different regions.

To clarify this point, the energy momentum tensor for an $\varepsilon = p$ fluid has the form

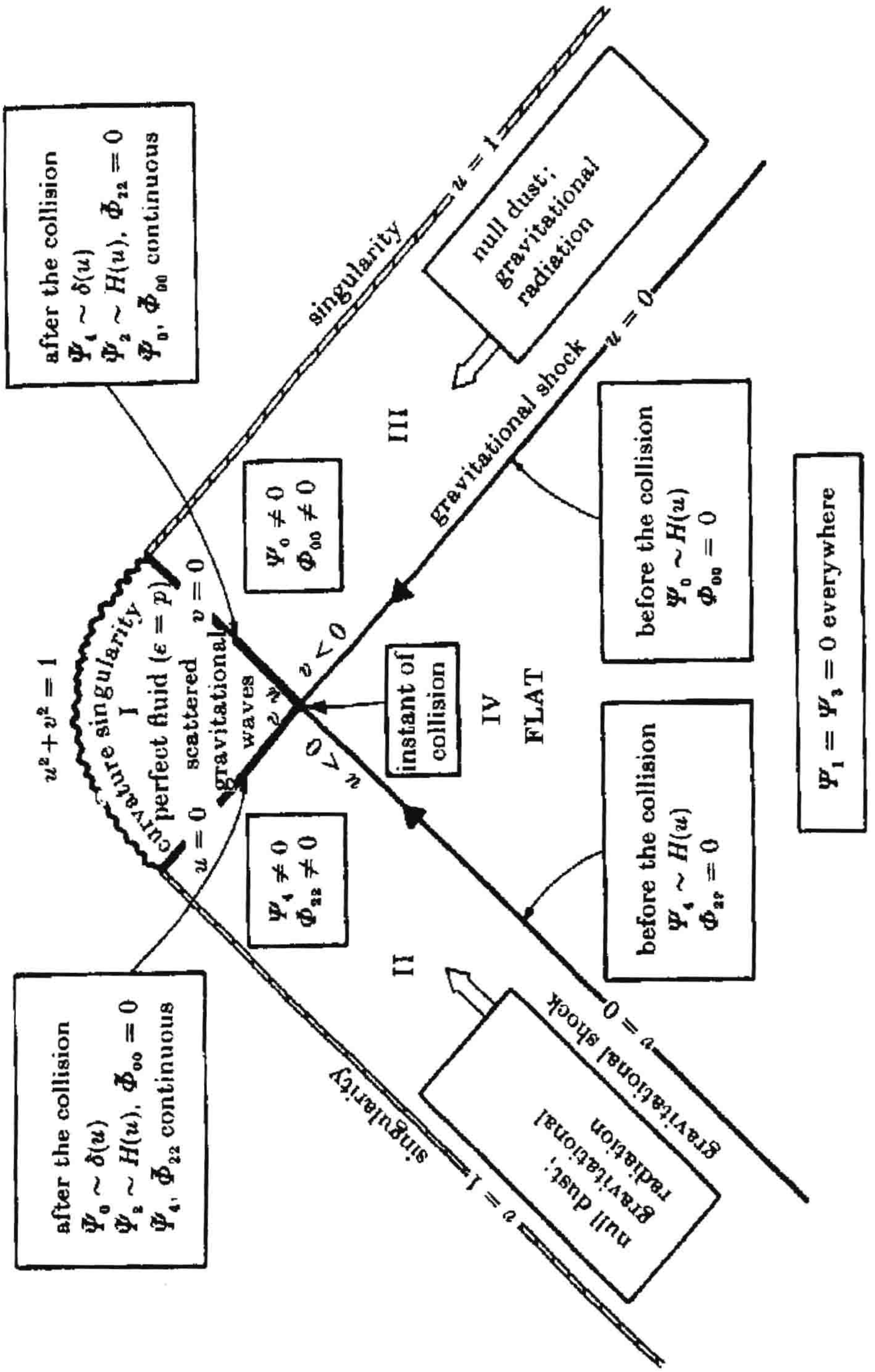
$$T^{ab} = 2v^a v^b - g^{ab}\left(g_{cd}v^c v^d\right):$$

where the vector v^a is directed along the fluid's 4-velocity and satisfies

$$v^{-}v_a = \varepsilon = p.$$

When v^a becomes a null vector, we get $\varepsilon = p = 0$, but this is consistent, and the energy-momentum tensor becomes $T \quad = 2v\,v$, namely null dust. Thus, null dust can be regarded as a particular (limiting) state of an $\varepsilon = p$ fluid.

In three subsequent papers (Chandrasekhar & Xanthopoulos 1986b, 1987a,b), Chandra confronted the issue of the singularity itself and the curiously slippery nature of the development of singularities in general relativity. The first of these papers presented a result which was a genuine surprise to me when Chandra first informed me about it. I had been under the impression (though without proof) that the circular line marked 'curvature singularity' at the top of the figure would always remain a region of infinite curvature for all collisions of this type. However,


after the collision
$\Psi_4 \sim \delta(u)$
$\Psi_2 \sim H(u)$, $\Phi_{22} = 0$
Ψ_0, Φ_{00} continuous
singularity
$u = 1$
null dust; gravitational radiation
$u = 0$
gravitational shock
III
$\Psi_0 \neq 0$
$\Phi_{00} \neq 0$
before the collision
$\Psi_0 \sim H(u)$
$\Phi_{00} = 0$
$u^2 + v^2 = 1$
curvature singularity
I
perfect fluid ($\varepsilon = p$)
scattered gravitational waves
$u = 0$
$v = 0$
$v < 0$
$u \geq$
$v \geq$
$u < 0$
instant of collision
IV
FLAT
$\Psi_1 = \Psi_3 = 0$ everywhere
$\Psi_4 \neq 0$
$\Phi_{22} \neq 0$
before the collision
$\Psi_4 \sim H(u)$
$\Phi_{22} = 0$
II
after the collision
$\Psi_0 \sim \delta(u)$
$\Psi_2 \sim H(u)$, $\Phi_{00} = 0$
Ψ_4, Φ_{22} continuous
singularity
$v = 1$
null dust; gravitational radiation
$v = 0$
gravitational shock

Chandra told me of his work in progress with Xanthopoulos on a vacuum colliding wave spacetime, providing an example which, at that stage, could be seen to have non-diverging curvature there. Soon after discussing the matter with me, Chandra was able to see how to change his coordinates so as to continue the metric explicitly across that seemingly singular region, exhibiting it to have more subtle geometric character than I had imagined, with a null hypersurface appearing that plays a role similar to that of an event horizon, and beyond which lies a timelike singularity. The remaining two coordinates cannot be ignored in this discussion, and the structure of the entire space-time remains somewhat complicated. In the other two papers, analogous situations are considered in which an electromagnetic field or an $\varepsilon = p$ fluid is introduced. The curiously contrasting behaviours are examined in some detail. Chandra's work with Xanthopoulos was clearly a high point for him, and it was a particular tragedy when Xanthopoulos met his untimely death from the attack of a crazed assassin. Chandra wrote a moving tribute in his *Selected Papers, Volume 6* (the volume that Xanthopoulos edited). Later he played an important role in the establishment of the tri-annual Xanthopoulos prize, for work in relativity achieved by young researchers.

4. Aspects of Chandra's mathematical heritage

The papers that I have discussed above far from exhaust Chandra's voluminous research into general relativity. There are several other articles in which various aspects of exact solutions are discussed. It is striking how his involvement with general relativity seduced him more and more in the direction of mathematics, where his earlier requirements of direct relevance to astrophysics seem to have been somewhat pushed aside. It seems clear that the equations that he encountered in his work on general relativity theory gave him immense satisfaction, and he seemed continually surprised at the mysterious beauty that these equations revealed. In his book *Truth and Beauty: Aesthetics and Motivations in Science* (Chandrasekhar 1987), he remarked on this explicitly, and in his booklet entitled *The Series Paintings of Claude Monet and the Landscape of General Relativity*, his dedication lecture on the opening of the Centre for Astronomy and Astrophysics in Pune (Chandrasekhar 1992), he compared some of this mathematical beauty with a sequence of paintings by Monet. Particular instances of the vacuum equations that had impressed him so much had occurred in a comparison that he had found between the basic black-hole space-times (Schwarzschild, Kerr) and the corresponding basic colliding wave space-times (Khan-Penrose, Nutku-Halil). A similar comparison was pointed out for the Einstein-Maxwell equations. All this arose from the mathematical structure of solutions of the Ernst equation (Ernst 1968), which controls the solutions of Einstein vacuum (or Einstein-Maxwell) metrics with two commuting symmetries.

As far as I am aware, these comparisons have not yet been fully understood in terms of the various techniques that have been developed over the years to handle such space-times. The particular procedure that appeals to me the most is that developed by Woodhouse and Mason (1988), which uses the procedures of twistor theory to describe these space-times in terms of Holomorphic vector bundles over non-Hausdorff Riemann spaces. This is part of a general comprehensive treatment of integrable systems (of which space-time metrics with two commuting Killing vector provide an example), in terms of twistor theory, which these two authors have thoroughly developed in a recent book (Mason and Woodhouse 1996). In view of the many intriguing relationships with different features of integral systems that Chandra's work has thrown up, I feel sure that there is a good deal that is deep, yet to be learned, from a study of the insights that he gained from his work in general relativity.

The same can also be said of a study of his analysis of the separation of gravitational perturbations and of other systems of equations in stationary black-hole backgrounds. There is yet much mystery to be unravelled. Some of this has already been achieved in the work of Carter (1968), Walker and Penrose (1970), Carter and McLenaghan (1979), Kamran and McLenaghan (1984) and many others, whereby separation can be related to the existence of a Killing Tensor, Killing spinor and Killing-Yano tensor. There are relations to twistor theory here also, and it is my guess that a further study of Chandra's work from this direction may well throw some profound light on these issues.

In a sense, Chandra's lifetime work was like a circle, basically starting with his insights that led us to believe that too massive white dwarfs must collapse to a space-time singularity, and finally reaching back to a sophisticated study of those very singularities. Yet, for all Chandra's extraordinary ingenuity and industry, the deep answers are still missing. I rather believe that Chandra did not really expect that this work might directly find an answer, however. The main thrust of his work in this area was somewhat different. It was driven more and more by the quest for mathematical elegance, coupled with a deep belief in a profound underlying connection between physics and mathematics. Perhaps it is here that the answers are finally to be found.

References

Bardeen, J. M., Press, W. H. 1973, *J. Math. Phys.*, **14**, 7.

Belinskii,V.A.,Khalatnikov,I.M., Lifshitz,E.M.1970, *Adv.Phys.*,**19**,523.

Belinskii,V.A., Khalatnikov, I.M., Lifshitz, E. M.1972, *Soviet Phys. JETP*, **62**, 1606.

Carter, B. 1968 "Hamilton–Jacobi and Schrodinger separable solutions of Einstein's equations", *Commun. Math. Phys.*, **10**, 280.

Carter, B. 1970. "An axisymmetric black hole has only two degrees of freedom", *Phys. Rev. Lett.*, **26**, 331.

Carter, B., McLenaghan, R. G. 1979, "Generalized total angular momentum operator for the Dirac equation in curved spacetime", *Phys. Rev. D.*, **19**, 1093.

Chandrasekhar,S.1931, "The maximum mass of ideal white dwarfs", *Astrophys.J.*,**74**,81.

Chandrasekhar, S. 1975, "On the Equations Governing the Perturbations of the Schwarzschild Black Hole", *Proc. R. Soc. London*, **A343**, 289.

Chandrasekhar, S., Detweiler, S. 1975a, "The Quasi-Normal Modes of the Schwarzschild Black Hole", *Proc. R. Soc. London*, **A344,** 441.

Chandrasekhar, S., Detweiler, S. 1975b, "On the Equations Governing the Axisymmetric Perturbations of the Kerr Black Hole", *Proc. R. Soc. London*, **A345,** 145.

Chandrasekhar, S. 1976a, "On a Transformation of Teukolsky's Equation and the Electromagnetic Perturbations of the Kerr Black Hole", *Proc. R. Soc. London*, **A348,** 39.

Chandrasekhar, S. 1976b, "The Solution of Maxwell's Equations in Kerr Geometry", *Proc. R.Soc. London*, **A349, 57**1.

Chandrasekhar, S. 1976c, "The Solution of Dirac's Equation in Kerr Geometry", *Proc. R. Soc. London*, **A349,** 571.

Chandrasekhar, S. 1978a, "The Kerr Metric and Stationary Axisymmetric Gravitational Fields", *Proc. R. Soc. London*, **A358,** 405.

Chandrasekhar, S. 1978b, "The Gravitational Perturbations of the Kerr Black Hole. I." *Proc. R. Soc. London*, **A358,** 421.

Chandrasekhar, S. 1978c, "The Gravitational Perturbations of the Kerr Black Hole. II." *Proc. R. Soc. London*, **A358,** 441.

Chandrasekhar, S. 1979a, "The Gravitational Perturbations of the Kerr Black Hole. III." *Proc. R. Soc. London*, **A365,** 425.

Chandrasekhar, S. 1979b, "On the Equations Governing the Perturbations of the Reissner-Nordström Black Hole", *Proc. R. Soc. London*, **A365,** 453.

Chandrasekhar, S., Xanthopoulos, B. C. 1979, "On the Metric Perturbations of the Reissner-Nordström Black Hole", *Proc. R. Soc. London*, **A367,** 1.

Chandrasekhar, S. 1980, "The Gravitational Perturbations of the Kerr Black Hole. IV. The Completion of the Solution", *Proc. R. Soc. London*, **A372,** 475.

Chandrasekhar, S. 1982, "On the Potential Barriers Surrounding the Schwarzschild Black Hole", *Spacetime and Geometry: The Alfred Schild Lectures,* ed. R. A. Matzner & L. C. Shepley, (Sheply, Austin: University of Texas Press).

Chandrasekhar, S., Hartle, J. B. 1982, "On Crossing the Cauchy Horizon of a Reissner Nordström Black Hole", *Proc. R. Soc. London*, **A384,** 301.

Chandrasekhar, S. 1983, *The Mathematical Theory of Black Holes*, (Oxford; Clarendon Press).

Chandrasekhar, S. 1984, "On Algebraically Special Perturbations of Black Holes", *Proc. R. Soc. London*, **A392,** 1.

Chandrasekhar, S., Ferrari, V. 1984, "On the Nutku-Halil Solution for Colliding Impulsive Gravitational Waves", *Proc. R. Soc. London*, **A396,** 55.

Chandrasekhar, S., Xanthopoulos, B. C. 1985a, "On Colliding Waves in the Einstein-Maxwell Theory", *Proc. R. Soc. London*, **A398**, 223.

Chandrasekhar, S., Xanthopoulos, B. C. 1985b, "On the Collision of Impulsive Gravitational Waves when Coupled with Fluid Motions" *Proc. R. Soc. London*, **A402**, 37.

Chandrasekhar, S., Xanthopoulos, B. C. 1986a, "On the Collision of Impulsive Gravitational Waves when Coupled with Null Dust", *Proc. R. Soc. London*, **A403**, 189.

Chandrasekhar, S., Xanthopoulos, B. C. 1986b, "A New Type of Singularity Created by Colliding Gravitational Waves", *Proc. R. Soc. London*, **A408**, 175.

Chandrasekhar, S. 1987 *Truth and Beauty: Aesthetics and Motivations in Science* (The University of Chicago Press).

Chandrasekhar, S., Xanthopoulos, B. C. 1987a, "On Colliding Waves that Develop Time-like Singularities", *Proc. R. Soc. London*, **A410**, 311.

Chandrasekhar, S., Xanthopoulos, B. C. 1987b, "The Effect of Sources on Horizons that may Develop when Plane Gravitational Waves Collide", *Proc. R. Soc. London*, **A414,** 1.

Chandrasekhar, S. 1989, "The Two-Centre Problem in General Relativity: The Scattering of Radiation by Two Extreme Reissner-Nordström Black Holes", *Proc. R. Soc. London,* **A421,** 227.

Chandrasekhar, S., Xanthopoulos B. C. 1989, "Two Black Holes Attached to Strings", *Proc. R. Soc.London*, **A423**, 387.

Chandrasekhar, S. 1990, "The Teukolsky-Starobinsky Constant for Arbitrary Spin", *Proc. R. Soc.*, **A430**, 433.

Chandrasekhar, S., Ferrari, V. 1990, "The Flux Integral for Axisymmetric Perturbations of Static Space-Times", *Proc. R. Soc.*, **A428**, 325.

Chandrasekhar, S. 1992, *The Series Paintings of Claude Monet and the Landscape of General Relativity*, (Pune: Inter-University Centre for Astronomy and Astrophysics).

Clarke, C. J.S. 1993,"The Analysis of Space-Time Singularities", *Cambridge Lecture Notes in Physics*, (Cambridge: Cambridge University Press).

Eddington, A. S. 1924, "A comparison of Whitehead's and Einstein's formulas", *Nature*, **113**, 192.

Eddington, A. S. 1946, *Fundamental Theory*, (Cambridge: Cambridge University Press).

Ernst, F. J. 1968, *Phys. Rev.*, **168,** 1415–7.

Geroch, R., Held, A., Penrose, R. 1973, "A space-time calculus based on pairs of null directions, *J. Math. Phys.*, **14**, 874–81.

Hawking, S. W. 1966, "The occurrence of singularities in cosmology", *Proc. R. Soc. (London),* **A294**, 511.

Hawking, S. W. 1972, "Black holes in general relativity", *Comm. Math. Phys.*, **25**, 152.

Hawking, S. W., Ellis, G. F. R. 1973, *The Large-Scale Structure of Space-Time,* (Cambridge: Cambridge University Press).

Hawking, S. W., Penrose, R. 1970, The singularities of gravitational collapse and cosmology, *Proc. R. Soc. (London)*, **A314**, 529-548.

Israel, W. 1967, "Event horizons in static vacuum space-times", *Phys. Rev.*, **164**, 1776.

Kamran, N., McLenaghan, R. G. 1984, "Separation of variables and symmetry operators for the neutrino and Dirac equations in space-times admitting a two-parameter abelian orthogonally transitive isometry group and a pair of shearfree geodesic null c Ongruences", *J. Math. Phys.*, **25**, 1019.

Kerr, R. P. 1963, "Gravitational field of a spinning mass as an example of algebraically special metrics", *Phys. Rev. Lett.*, **11**, 237.

Khan, K. A., Penrose, R. 1971, *Nature*, **229**,185.

Landau, L. 1932, "On the theory of stars", *Phys. Z. Sowjetunion I*, **285**.

Lemaîitre, G. 1933, "L'universe en expansion", *Ann. Soc. Sci. Bruxelles I*, **A53**, 51. (cf. p.82).

Lifshitz, E. M., Khalatnikov, I. M. 1963, "Investigations in relativistic cosmology", in *Advances in Physics*, **12**, 185.

Mason, L. J., Woodhouse, N. M. J. 1996, *Integrability, Self-Duality, and Twistor Theory*, (Oxford: Oxford University Press).

McNamara, J. M. 1978a, *Proc. R. Soc. London*, **A358**, 499.

McNamara, J. M. 1978b, *Proc. R. Soc. London*, **A364**, 121.

Newman, E. T., Couch, E., Chinnapared, K., Exton, A., Prakash, A., Torrence, R. 1965, "Metric of a rotating charged mass", *J. Math. Phys.*, **6**, 918–9.

Newman, E. T., Penrose, R. 1962, "An approach to gravitational radiation by a method of spin coefficients", *J. Math. Phys.*, **3**, 896 (Errata **4**, 1963, 998).

Nutku, Y., Halil, M. 1977, *Phys. Rev. Lett.*, **39**, 1379.

Oppenheimer, J. R., Snyder, H. 1939, "On continued gravitational contraction", *Phys. Rev.*, **56**, 455–9.

Oppenheimer, J. R., Volkoff, H. 1939, "On massive neutron cores", *Phys. Rev.*, **55**, 374.

Page, D. N. 1976, "Dirac equation around a charged, rotating black hole", *Phys. Rev. D.*, **14**, 1509.

Penrose, R. 1965, "Gravitational collapse and space-time singularities", *Phys. Rev. Lett.*, **14**, 579.

Penrose, R. 1969, "Gravitational collapse: the role of general relativity, *Rivista del Nuovo Cimento*", *Numero speciale,* **1**, 252–276.

Penrose, R. 1973, *Ann. N.Y. Acad. Sci.*, **224**, 125.

Penrose, R. 1978, "Singularities of Space-Time", in *Theoretical Principles in Astrophysics and Relativity,* eds. N. R. Lebowitz, W. H. Reid & P. O. Vandervoort (Chicago: Chicago University Press, Chicago).

Penrose, R. 1979, "Singularities and Time-Asymmetry", in *General Relativity,* eds. S. W. Hawking & W. Israel (Cambridge: Cambridge University Press).

Robinson, D. C. 1975, *Phys. Rev. Lett.*, **34**, 905.

Regge, T., Wheeler, J. A. W., 1957, "Stability of a Schwarzschild singularity", *Phys. Rev.*, **108**, 1063.

Simpson, M., Penrose, R. 1973, *Int. J. Theor. Phys.*, **7**, 183.

Synge, J. L. 1950, "The gravitational field of a particle", *Proc. Irish Acad.*, **A53**, 83.

Teukolsky, S. A. 1973, "Perturbations of a rotating black hole: Fundamental equations for gravitational, electromagnetic, and neutrino-field perturbations", *Astrophys. J.*, **185**, 635.

Walker, M., Penrose, R. 1970, "On quadratic first integrals of the geodesic equations for type 22 spacetimes", *Comm. Math. Phys.*, **18**, 265.

Woodhouse, N. M. J., Mason, L. J. 1988, "The Geroch group and non-Hausdorff twistor spaces", *Nonlinearity*, **1**, 73–114.

Zerilli, F. J. 1970, *Phys. Rev. D.*, **2**, 2141.

246

CONSCIOUS EVENTS AS ORCHESTRATED SPACE-TIME SELECTIONS

J. Consc. Stud. **3** (1996), 36–53. With S. Hameroff. Also published in *Explaining Consciousness – The 'Hard Problem'*. Ed. J. Shear, 177–195, MIT Press, 1997

This is my first collaborative publication with Stuart Hameroff. After reading *The Emperor's New Mind* (Oxford University Press, 1989) he wrote to me, telling me about microtubules and suggesting that they might provide the key missing biological ingredient to the ideas I had put forward in my book. We then got together and eventually formulated the general scheme described here. It is natural that many people, knowledgeable about the kind of circumstance normally needed for collective quantum effects such as superconductivity, would express scepticism about our suggested role for (A-lattice) body-temperature microtubules. However, some very recent experiments (as of May 2010) by Satyajit Sahu, Anirban Bandyopadhyay, and Daisuke Fujita seem to indicate some very remarkable effects of this actual nature.

662 Conscious events as orchestrated space-time selections. *J. Consc. Stud.* **3** (1996), 36–53.
With S. Hameroff. Also published in *Explaining Consciousness – The 'Hard Problem'*.
Ed. J. Shear, 177–195, 1997

Stuart R. Hameroff and Roger Penrose

Conscious Events as Orchestrated Space–Time Selections

Introduction: Self-Selection in an Experiential Medium?

The 'hard problem' of incorporating the phenomenon of consciousness into a scientific world-view involves finding scientific explanations of qualia, or the subjective experience of mental states (Chalmers, keynote paper and 1996). On this, reductionist science is still at sea. Why do we have an inner life, and what exactly is it?

One set of philosophical positions, addressing the hard problem, views consciousness as a fundamental component of physical reality. For example an extreme view — 'panpsychism' — is that consciousness is a quality of all matter: atoms and their subatomic components having elements of consciousness (e.g. Spinoza, 1677; Rensch, 1960). 'Mentalists' such as Leibniz and Whitehead (e.g. 1929) contended that systems ordinarily considered to be physical are constructed in some sense from mental entities. Bertrand Russell (1954) described 'neutral monism' in which a common underlying entity, neither physical nor mental, gave rise to both. Recently Stubenberg (1996) has claimed that qualia are that common entity. In monistic idealism, matter and mind arise from consciousness — the fundamental constituent of reality (e.g. Goswami, 1993). Wheeler (1990) has suggested that information is fundamental to the physics of the universe. From this, Chalmers (keynote paper and 1996) proposes a double-aspect theory in which information has both physical and experiential aspects.

Among these positions, the philosophy of Alfred North Whitehead (1929; 1933) may be most directly applicable. Whitehead describes the ultimate concrete entities in the cosmos as being actual 'occasions of experience', each bearing a quality akin to 'feeling'. Whitehead construes 'experience' broadly — in a manner consistent with panpsychism — so that even 'temporal events in the career of an electron have a kind of 'protomentality'. Whitehead's view may be considered to differ from panpsychism, however, in that his discrete 'occasions of experience' can be taken to be related to 'quantum events' (Shimony, 1993). In the standard descriptions of quantum mechanics, randomness occurs in the events described as quantum state reductions — these being events which appear to take place when a quantum-level process gets magnified to a macroscopic scale.

This paper was originally published in the *Journal of Consciousness Studies*, **3**, No.1 (1996), pp. 36–53.

Quantum state reduction (here denoted by the letter **R**; cf. Penrose 1989, 1994) is the random procedure that is adopted by physicists in their descriptions of the quantum measurement process. It is still a highly controversial matter whether **R** is to be taken as a 'real' physical process, or whether it is some kind of illusion and not to be regarded as a fundamental ingredient of the behaviour of Nature. Our position is to take **R** to be indeed real — or, rather to regard it as a close approximation to an objectively real process **OR** (objective reduction), which is to be a non-computable process instead of merely a random one (see Penrose 1989; 1994). In almost all physical situations, **OR** would come about in situations in which the random effects of the environment dominate, so **OR** would be virtually indistinguishable from the random **R** procedure that is normally adopted by quantum theorists. However, when the quantum system under consideration remains coherent and well isolated from its environment, then it becomes possible for its state to collapse spontaneously, in accordance with the **OR** scheme we adopt, and to behave in non-computable rather than random ways. Moreover, this **OR** scheme intimately involves the geometry of the physical universe at its deepest levels.

Our viewpoint is to regard experiential phenomena as also inseparable from the physical universe, and in fact to be deeply connected with the very laws which govern the physical universe. The connection is so deep, however, that we perceive only glimmerings of it in our present day physics. One of these glimmerings, we contend, is a necessary non-computability in conscious thought processes; and we argue that this non-computability must also be inherent in the phenomenon of quantum state *self*-reduction — the 'objective reduction' (**OR**) referred to above. This is the main thread of argument in *Shadows of the Mind* (Penrose, 1994). The argument that conscious thought, whatever other attributes it may also have, is non-computable (as follows most powerfully from certain deductions from Gödel's incompleteness theorem) grabs hold of one tiny but extremely valuable point. This means that at least some conscious states cannot be derived from previous states by an algorithmic process — a property which distinguishes human (and other animal) minds from computers. Noncomputability *per se* does not directly address the 'hard problem' of the nature of experience, but it is a clue to the kind of physical activity that lies behind it. This points to **OR**, an underlying physical action of a completely different character from that which would appear to underlie nonconscious activity. Following this clue with sensitivity and patience should ultimately lead to real progress towards understanding mental phenomena in their inward manifestations as well as outward.

In the **OR** description, consciousness takes place if an organized quantum system is able to isolate and sustain coherent superpositions until its quantum gravity threshold for space–time separation is met; it then *self*-reduces (non-computably). For consciousness to occur, *self*-reduction is essential, as opposed to the reduction being triggered by the system's random environment. (In the latter case, the reduction would itself be effectively random and would lack useful non-computability, being unsuitable for direct involvement in consciousness.) We take the *self*-reduction to be an instantaneous event — the climax of a *self*-organizing process fundamental to the structuring of space–time — and apparently consistent with a Whitehead 'occasion of experience'.

As **OR** could, in principle, occur ubiquitously within many types of inanimate media, it may seem to imply a form of 'panpsychism' (in which individual electrons, for example, possess an experiential quality). However according to the principles of **OR** (as expounded in Penrose, 1994; 1996), a single superposed electron would spontaneously reduce its state (assuming it could maintain isolation) only once in a period longer than the present age of the universe. Only large collections of particles acting coherently in a single macroscopic quantum state could possibly sustain isolation and support coherent superposition in a timeframe brief enough to be relevant to our consciousness. Thus, only very special circumstances could support consciousness:

1) High degree of coherence of a quantum state — a collective mass of particles in superposition for a time period long enough to reach threshold, and brief enough to be useful in thought processes.
2) Ability for the **OR** process to be at least transiently isolated from a 'noisy' environment until the spontaneous state reduction takes place. This isolation is required so that reduction is not simply random. Mass movement in the environment which entangles with the quantum state would effect a random (not non-computable) reduction.
3) Cascades of **OR**s to give a 'stream' of consciousness, huge numbers of **OR** events taking place during the course of a lifetime.

By reaching the quantum gravity threshold, each **OR** event has a fundamental bearing on space–time geometry. One could say that a cascade of **OR** events charts an actual course of physical space–time geometry selections.

It may seem surprising that quantum gravity effects could plausibly have relevance at the physical scales relevant to brain processes. For quantum gravity is normally viewed as having only absurdly tiny influences at ordinary dimensions. However, we shall show later that this is not the case, and the scales determined by basic quantum gravity principles are indeed those that are relevant for conscious brain processes.

We must ask how such an **OR** process could actually occur in the brain? How could it be coupled to neural activities at a high rate of information exchange; how could it account for preconscious to conscious transitions, have spatial and temporal binding, and both simultaneity and time flow?

We here nominate an **OR** process with the requisite characteristics to be occurring in cytoskeletal microtubules within the brain's neurons. In our model, microtubule-associated proteins 'tune' the quantum oscillations leading to **OR**; we thus term the process 'orchestrated objective reduction' (**Orch OR**).

Space–Time: Quantum Theory and Einstein's Gravity

Quantum theory describes the extraordinary behaviour of the matter and energy which comprise our universe at a fundamental level. At the root of quantum theory is the wave/particle duality of atoms, molecules, and their constituent particles. A quantum system such as an atom or subatomic particle which remains isolated from its environment behaves as a 'wave of possibilities' and exists in a coherent complex-number valued 'superposition' of many possible states. The behaviour of such wave-like, quantum-level objects can be satisfactorily described in terms of a state vector which evolves deterministically according to the Schrödinger equation (unitary evolution), denoted by **U**.

Conscious events as orchestrated space-time selections. *J. Consc. Stud.* **3** (1996), 36–53.
With S. Hameroff. Also published in *Explaining Consciousness – The 'Hard Problem'*.
Ed. J. Shear, 177–195, 1997

CHAPTER 246

Somehow, quantum microlevel superpositions lead to unsuperposed stable structures in our macro-world. In a transition known as wave function collapse, or state reduction (**R**), the quantum wave of alternative possibilities reduces to a single macroscopic reality — an 'eigenstate' of some appropriate operator. (This would be just one out of many possible alternative eigenstates relevant to the quantum operator.) This process is invoked in the description of a macroscopic measurement, when effects are magnified from the small, quantum scale to the large, classical scale.

According to conventional quantum theory (as part of the standard 'Copenhagen interpretation'), each choice of eigenstate is entirely random, weighted according to a probability value that can be calculated from the previous state according to the precise procedures of the quantum formalism. This probabilistic ingredient was a feature with which Einstein, among others, expressed displeasure: 'You believe in a God who plays dice, and I in complete law and order.'[1] Penrose (1989; 1994) has contended that, at a deeper level of description, the choices may more accurately arise as the result of some presently unknown 'non-computational' mathematical/physical (i.e. 'Platonic realm') theory, that is they cannot be deduced algorithmically. Penrose argues that such non-computability is essential to consciousness, because (at least some) conscious mental activity is unattainable by computers.

It can be argued that present-day physics has no clear explanation for the cause and occurrence of wave function collapse **R**. Experimental and theoretical evidence through the 1930's led quantum physicists (such as Schrödinger, Heisenberg, Dirac, von Neumann, and others) to postulate that quantum-coherent superpositions persist indefinitely in time, and would, in principle be maintained from the micro to macro levels. Or perhaps they would persist until conscious observation collapses, or reduces, the wave function (subjective reduction, or '**SR**'). Accordingly, even macroscopic objects, if unobserved, could remain superposed. To illustrate the apparent absurdity of this notion, Erwin Schrödinger (e.g. 1935) described his now-famous 'cat in a box' being simultaneously *both* dead and alive until the box was opened and the cat observed.

As a counter to this unsettling prospect, various new physical schemes for collapse according to objective criteria (objective reduction — '**OR**') have recently been proposed. According to such a scheme, the growth and persistence of superposed states could reach a critical threshold, at which collapse, or **OR** rapidly occurs (e.g. Pearle, 1989; Ghirardi *et al.*, 1986). Some of these schemes are based specifically on gravitational effects mediating **OR** (e.g. Károlyházy, 1966; Károlyházy *et al.*, 1986; Diósi, 1989; Ghirardi *et al.*, 1990; Penrose, 1989; 1994; Pearle and Squires, 1995; Percival, 1995). Table 1 categorizes types of reduction.

The physical phenomenon of gravity, described to a high degree of accuracy by Isaac Newton's mathematics in 1687, has played a key role in scientific understanding. However, in 1915, Einstein created a major revolution in our scientific world-view. According to Einstein's theory, gravity plays a unique role in physics for several reasons (cf. Penrose, 1994). Most particularly, these are:

1) Gravity is the only physical quality which influences causal relationships between space–time events.

---

[1] From a letter to Max Born.

2) Gravitational force has no local reality, as it can be eliminated by a change in space–time coordinates; instead, gravitational tidal effects provide a *curvature* for the very *space–time* in which all other particles and forces are contained.

Context	**Cause of Collapse (Reduction)**	**Description**	**Acronym**
Quantum coherent superposition	No collapse	Evolution of wave function (Schrödinger equation)	**U**
Conventional quantum theory (Copenhagen interpretation)	Environmental entanglement, Measurement, Conscious observation	Reduction; Subjective reduction	**R** **SR**
New physics (Penrose,1994)	Self-collapse — quantum gravity induced (Penrose, Diosi, etc.)	Objective reduction	**OR**
Consciousness (present paper)	Self-collapse, quantum gravity threshold in microtubules orchestrated by MAPs etc.	Orchestrated objective reduction	**Orch OR**

Table 1: Description of wave function collapse

It follows from this that gravity cannot be regarded as some kind of 'emergent phenomenon', secondary to other physical effects, but must be a 'fundamental component' of physical reality.

There are strong arguments (e.g. Penrose, 1987; 1996) to suggest that the appropriate union of general relativity (Einstein's theory of gravity) with quantum mechanics — a union often referred to as 'quantum gravity' — will lead to a significant change in *both* quantum theory and general relativity, and, when the correct theory is found, will yield a profoundly *new* understanding of physical reality. And although gravitational *forces* between objects are exceedingly weak (feebler than, for example, electrical forces by some 40 orders of magnitude), there are significant reasons for believing that gravity has a fundamental influence on the behaviour of quantum systems as they evolve from the micro to the macrolevels. The appropriate union of quantum gravity with biology, or at least with advanced biological nervous systems, may yield a profoundly new understanding of consciousness.

Curved Space–Time Superpositions and Objective Reduction ('OR')

According to modern accepted physical pictures, reality is rooted in 3-dimensional space and a 1-dimensional time, combined together into a 4-dimensional space–time. This space–time is slightly curved, in accordance with Einstein's general theory of relativity, in a way which encodes the gravitational fields of all distributions of mass density. Each mass density effects a space–time curvature, albeit tiny.

This is the standard picture according to *classical* physics. On the other hand, when *quantum* systems have been considered by physicists, this mass-induced tiny curvature in the structure of space–time has been almost invariably ignored, gravitational effects having been assumed to be totally insignificant for normal problems in which quantum theory is important. Surprising as it may seem, however, such tiny differences in space–time structure *can* have large effects: for they entail subtle but fundamental influences on the very rules of quantum mechanics.

Superposed quantum states for which the respective mass distributions differ significantly from one another will have space–time geometries which correspondingly differ. Thus, according to standard quantum theory, the superposed state would have to involve a quantum superposition of these differing space–times. In the absence of a coherent theory of quantum gravity there is no accepted way of handling such a superposition. Indeed, the basic principles of Einstein's general relativity begin to come into profound conflict with those of quantum mechanics (cf. Penrose, 1996). Nevertheless, various tentative procedures have been put forward in attempts to describe such a superposition. Of particular relevance to our present proposals are the suggestions of certain authors (e.g. Károlyházy, 1966; 1974; Károlyházy *et al.*, 1986; Kibble, 1981; Diósi, 1989; Ghirardi *et al.*, 1990; Pearle and Squires, 1995; Percival, 1995; Penrose, 1993; 1994; 1996) that it is at this point that an objective quantum state reduction (**OR**) ought to occur, and the rate or timescale of this process can be calculated from basic quantum gravity considerations. These particular proposals differ in certain detailed respects, and for definiteness we shall follow the specific suggestions made in Penrose (1994; 1996). Accordingly, the quantum superposition of significantly differing space–times is unstable, with a life-time given by that timescale. Such a superposed state will decay — or 'reduce' — into a single universe state, which is one or the other of the space–time geometries involved in that superposition.

Whereas such an **OR** action is not a generally recognized part of the normal quantum-mechanical procedures, there is no plausible or clearcut alternative that standard quantum theory has to offer. This **OR** procedure avoids the need for 'multiple universes' (cf. Everett, 1957; Wheeler, 1957, for example). There is no agreement, among quantum gravity experts, about how else to address this problem. For the purposes of the present article, it will be assumed that a gravitationally induced **OR** action is indeed the correct resolution of this fundamental conundrum.

Figure 1 (adapted from Penrose, 1994, p. 338) schematically illustrates the way in which space time structure can be affected when two macroscopically different mass distributions take part in a quantum superposition. Each mass distribution gives rise to a separate space–time, the two differing slightly in their curvatures. So long as the two distributions remain in quantum superposition, we must consider that the two space–times remain in superposition. Since, according to the principles of general relativity, there is no natural way to identify the points of one space–time with corresponding points of the other, we have to consider the two as separated from one another in some sense, resulting in a kind of 'blister' where the space–time bifurcates.

A bifurcating space–time is depicted in the lowest of the three diagrams, this being the union ('glued together version') of the two alternative space–time histories that are depicted at the top of Figure 1. The initial part of each space–time is at the lower end of

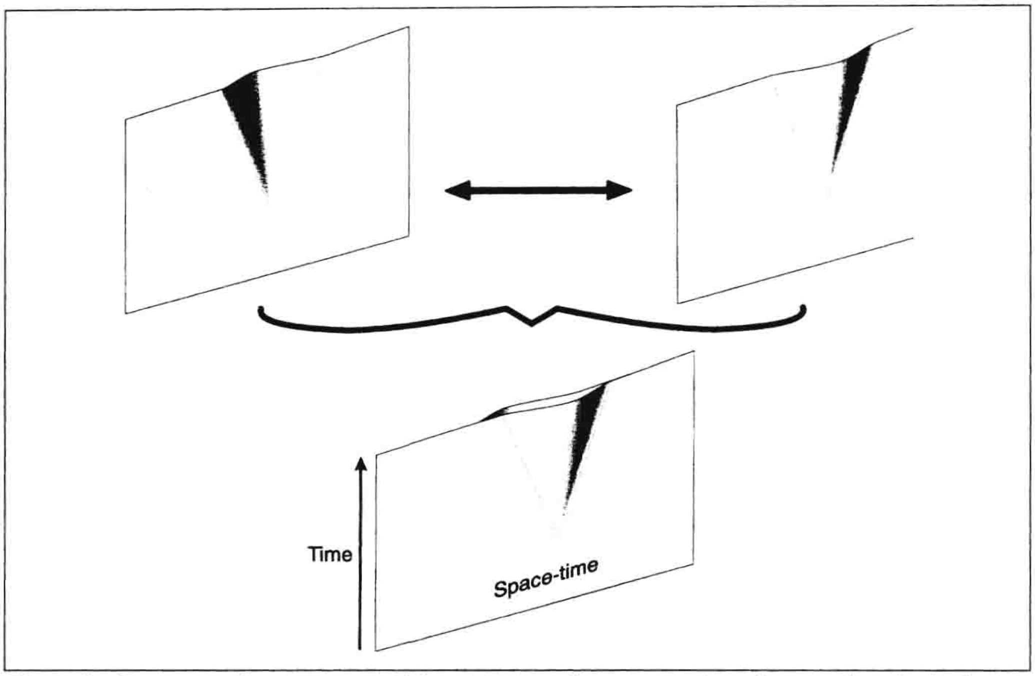


Figure 1. Quantum coherent superposition represented as a separation of space–time. In the lowest of the three diagrams, a bifurcating space–time is depicted as the union ('glued together version'), of the two alternative space–time histories that are depicted at the top of the Figure. The bifurcating space–time diagram illustrates two alternative mass distributions actually in quantum superposition, whereas the top two diagrams illustrate the two individual alternatives which take part in the superposition (adapted from Penrose, 1994, p. 338).

each individual space–time diagram. The bottom space–time diagram (the bifurcating one) illustrates two alternative mass distributions actually in quantum superposition, whereas the top two illustrate the two individual alternatives which take part in the superposition. The combined space–time describes a superposition in which the alternative locations of a mass move gradually away from each other as we proceed in the upward direction in the diagram. Quantum mechanically (i.e. so long as **OR** has not taken place), we must think of the 'physical reality' of this situation as being illustrated as an actual superposition of these two slightly differing space–time manifolds, as indicated in the bottom diagram. As soon as **OR** has occurred, one of the two individual space–times takes over, as depicted as one of the two sheets of the bifurcation. For clarity only, the bifurcating parts of these two sheets are illustrated as being one convex and the other concave. Of course there is additional artistic licence involved in drawing the space–time sheets as 2-dimensional, whereas the actual space–time constituents are 4-dimensional. Moreover, there is no significance to be attached to the imagined '3-dimensional space' within which the space–time sheets seem to be residing. There is no 'actual' higher dimensional space there, the 'intrinsic geometry' of the bifurcating space–time being all that has physical significance. When the 'separation' of the two space–time sheets reaches a critical amount, one of the two sheets 'dies' — in accordance with the **OR** criterion — the other being the one that persists in physical reality. The quantum state thus reduces (**OR**), by choosing between either the 'concave' or 'convex' space–time of Figure 1.

It should be made clear that this measure of separation is only very schematically illustrated as the 'distance' between the two sheets in the lower diagram in Figure 1. As remarked above, there is no physically existing 'ambient higher-dimensional space' inside which the two sheets reside. The degree of separation between the space–time sheets is a more abstract mathematical thing; it would be more appropriately described in terms of a *symplectic measure* on the space of 4-dimensional metrics (cf. Penrose, 1993) — but the details (and difficulties) of this will not be important for us here. It may be noted, however, that this separation is a space–time separation, not just a spatial one. Thus the *time* of separation contributes as well as the spatial displacement. Roughly speaking, it is the product of the temporal separation $\boldsymbol{T}$ with the spatial separation $\boldsymbol{S}$ that measures the overall degree of separation, and **OR** takes place when this overall separation reaches the critical amount. [This critical amount would be of the order of unity, in absolute units, for which the Planck-Dirac constant $\hbar$ $(= h/2\pi)$, the gravitational constant G, and the velocity of light c, all take the value unity; cf. Penrose, 1994, pp. 337–9.] Thus for small $\boldsymbol{S}$, the lifetime $\boldsymbol{T}$ of the superposed state will be large; on the other hand, if $\boldsymbol{S}$ is large,then $\boldsymbol{T}$ will be small. To calculate $\boldsymbol{S}$, we compute (in the Newtonian limit of weak gravitational fields) the gravitational self-energy $\boldsymbol{E}$ of the difference between the mass distributions of the two superposed states. (That is, one mass distribution counts positively and the other, negatively; see Penrose, 1994; 1996.) The quantity $\boldsymbol{S}$ is then given, in absolute units, by:

$$\boldsymbol{S} = \boldsymbol{E}$$

Thus, restoring standard units,

$$\boldsymbol{T} = \hbar\, \boldsymbol{E}^{-1}$$

Schematically, since $\boldsymbol{S}$ represents three dimensions of displacement rather than the one dimension involved in $\boldsymbol{T}$, we can imagine that this displacement is shared equally between each of these three dimensions of space — and this is what has been depicted in Figure 3 (below). However, it should be emphasized that this is for pictorial purposes only, the appropriate rule being the one given above. These two equations relate the mass distribution, time of coherence, and space–time separation for a given **OR** event. If, as some philosophers contend, experience is contained in space–time, **OR** events are *self-organizing* processes in that experiential medium, and a candidate for consciousness.

But where in the brain, and how, could coherent superposition and **OR** occur? A number of sites and various types of quantum interactions have been proposed. We strongly favour microtubules as an important ingredient; however various organelles and biomolecular structures including clathrins, myelin (glial cells), presynaptic vesicular grids (Beck and Eccles, 1992) and neural membrane proteins (Marshall, 1989) might also participate.

Microtubules

Properties of brain structures suitable for quantum coherent superposition, **OR** and relevant to consciousness might include: 1) high prevalence, 2) functional importance (for example regulating neural connectivity and synaptic function), 3) periodic, crystal-like lattice dipole structure with longrange order, 4) ability to be transiently isolated from external interaction/observation, 5) functionally coupled to quantum-level events, 6)

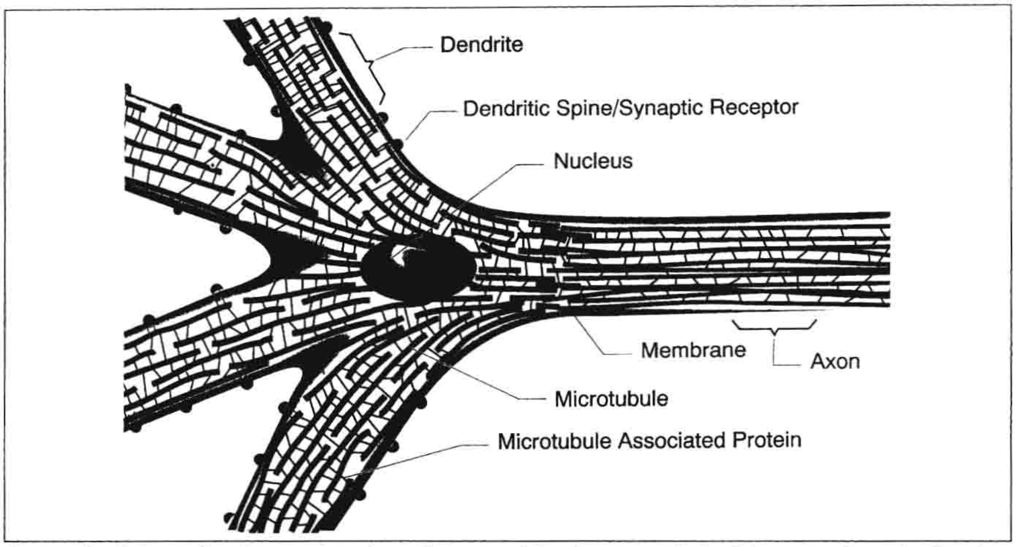

Dendrite
Dendritic Spine/Synaptic Receptor
Nucleus
Membrane
Axon
Microtubule
Microtubule Associated Protein

Figure 2. Schematic of central region of neuron (distal axon and dendrites not shown), showing parallel arrayed microtubules interconnected by MAPs. Microtubules in axons are lengthy and continuous, whereas in dendrites they are interrupted and of mixed polarity. Linking proteins connect microtubules to membrane proteins including receptors on dendritic spines.

hollow, cylindrical (possible wave-guide), and 7) suitable for information processing. Membranes, membrane proteins, synapses, DNA and other types of structures have some, but not all, of these characteristics. Cytoskeletal microtubules appear to qualify in all respects.

Interiors of living cells, including the brain's neurons, are spatially and dynamically organized by *self*-assembling protein networks: the cytoskeleton. Within neurons, the cytoskeleton establishes neuronal form, and maintains and regulates synaptic connections. Its major components are microtubules, hollow cylindrical polymers of individual proteins known as tubulin. Microtubules ('MTs') are interconnected by linking proteins (microtubule-associated proteins: 'MAPs') to other microtubules and cell structures to form cytoskeletal lattice networks (Figure 2).

MTs are hollow cylinders 25 nanometers (nm) in diameter whose lengths vary and may be quite long within some nerve axons. MT cylinder walls are comprised of 13 longitudinal protofilaments which are each a series of subunit proteins known as tubulin (Figure 3). Each tubulin subunit is a polar, 8 nm dimer which consists of two slightly different 4 nm monomers (α and β tubulin — Figure 4). Tubulin dimers are dipoles, with surplus negative charges localized toward monomers (DeBrabander, 1982), and within

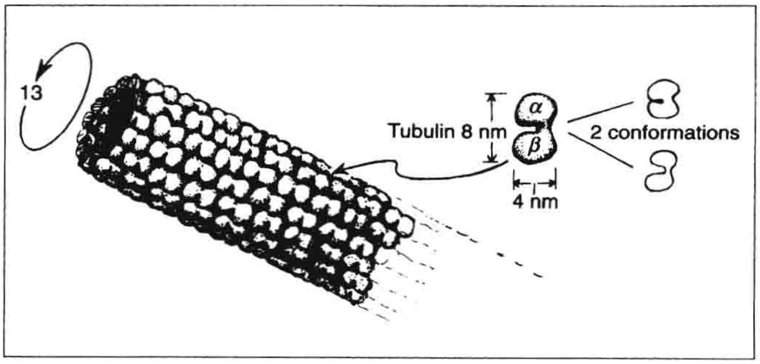


Figure 3.

Microtubule structure: a hollow tube of 25 nanometers diameter, consisting of 13 columns of tubulin dimers. Each tubulin molecule is capable of (at least), two conformations. (Reprinted with permission from Penrose, 1994, p. 359.)

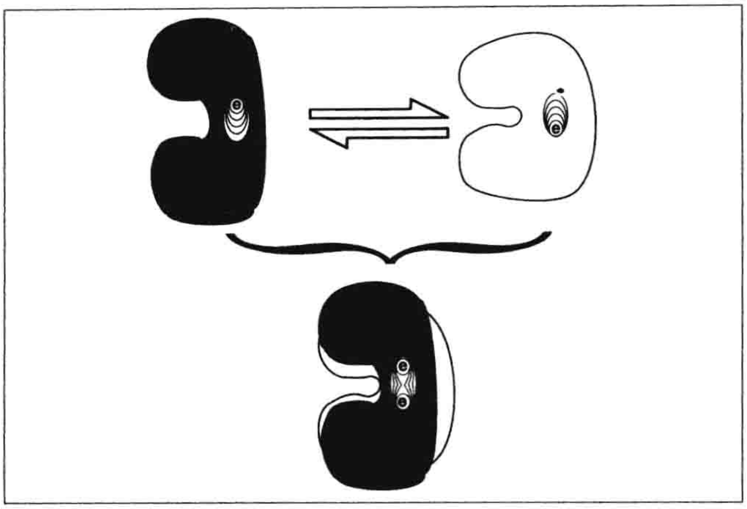


Figure 4.

Top:
Two states of tubulin in which a *single* quantum event (electron localization), within a central hydrophobic pocket is coupled to a *global* protein conformation. Switching between the two states can occur on the order of nanoseconds to picoseconds.

Bottom:
Tubulin in quantum coherent superposition.

MTs are arranged in a hexagonal lattice which is slightly twisted, resulting in helical pathways which repeat every 3, 5, 8 and other numbers of rows. Traditionally viewed as the cell's 'bonelike' scaffolding, microtubules and other cytoskeletal structures also appear to fill communicative and information processing roles. Numerous types of studies link the cytoskeleton to cognitive processes (for review, cf. Hameroff and Penrose, 1996). Theoretical models and simulations suggest how conformational states of tubulins within microtubule lattices can interact with neighbouring tubulins to represent, propagate and process information as in molecular-level 'cellular automata', or 'spinglass' type computing systems (Figure 5; e.g. Hameroff and Watt, 1982; Rasmussen *et al.*, 1990; Tuszyński *et al.*, 1995). In Hameroff and Penrose (1996, and in summary form Penrose and Hameroff, 1995), we present a model linking microtubules to consciousness, using quantum theory as viewed in the particular 'realistic' way referred to above, and as described in *Shadows of the Mind* (Penrose, 1994). In our model, quantum coherence emerges, and is isolated, in brain microtubules until the differences in mass–energy distribution among superposed tubulin states reach the above threshold of insta-

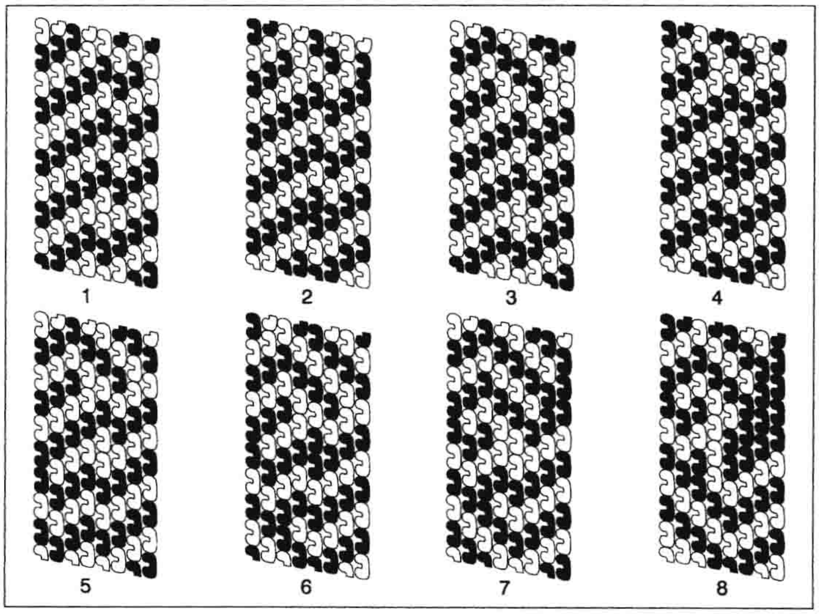


Figure 5. Microtubule automaton simulation (from Rasmussen *et al.*, 1990). Black and white tubulins correspond to states shown in Figure 2. Eight nanosecond time steps of a segment of one microtubule are shown in 'classical computing' mode in which patterns move, evolve, interact and lead to emergence of new patterns.

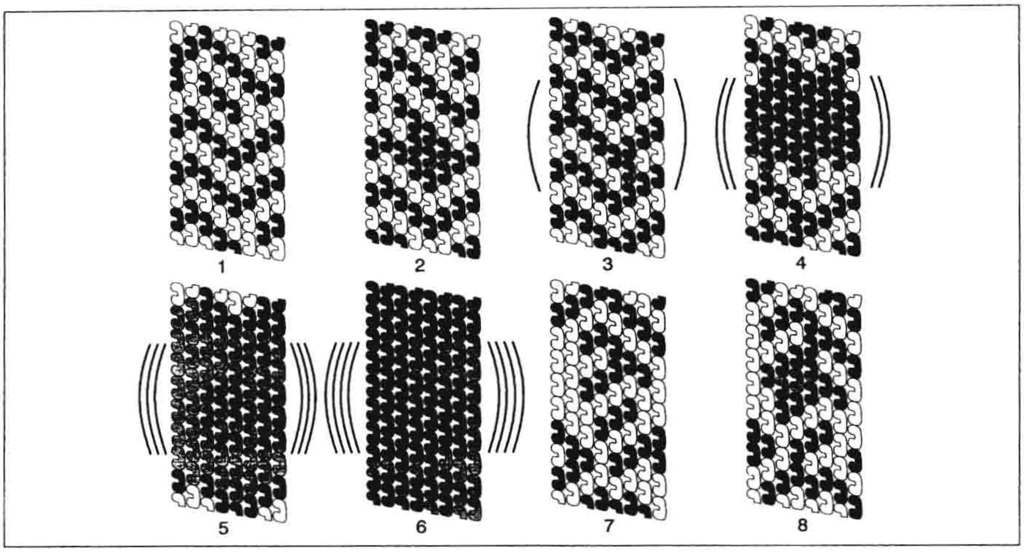


Figure 6. Microtubule automaton sequence simulation in which classical computing (step 1), leads to emergence of quantum coherent superposition (steps 2–6), in certain (grey) tubulins due to pattern resonance. Step 6 (in coherence with other microtubule tubulins) meets critical threshold related to quantum gravity for self–collapse (**Orch OR**). Consciousness (**Orch OR**) occurs in the step 6 to 7 transition. Step 7 represents the eigenstate of mass distribution of the collapse which evolves by classical computing automata to regulate neural function. Quantum coherence begins to re–emerge in step 8.

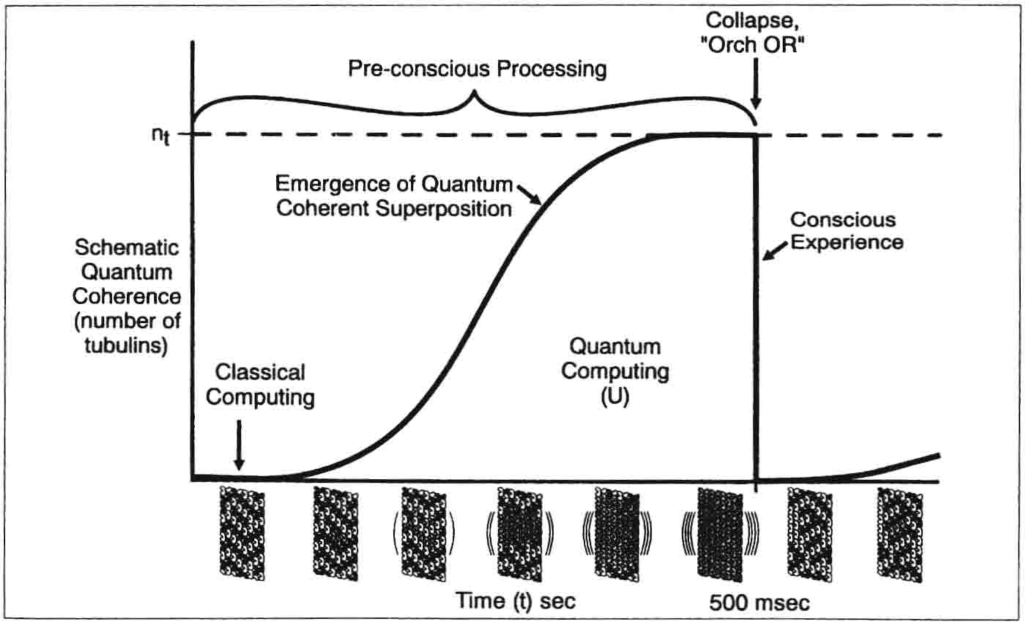


Figure 7. Schematic graph of proposed quantum coherence (number of tubulins), emerging vs time in microtubules. 500 milliseconds is time for pre–conscious processing (e.g. Libet *et al.*, 1979). Area under curve connects mass–energy differences with collapse time in accordance with gravitational **OR**. This degree of coherent superposition of differing space–time geometries leads to abrupt quantum → classical reduction ('self–collapse' or 'orchestrated objective reduction: **Orch OR**').

bility, related to quantum gravity (Figure 6). The resultant *self*-collapse (**OR**), considered to be a time-irreversible process, creates an instantaneous 'now' event in our model. Sequences of such events create a flow of time and consciousness (Figures 7 and 8).

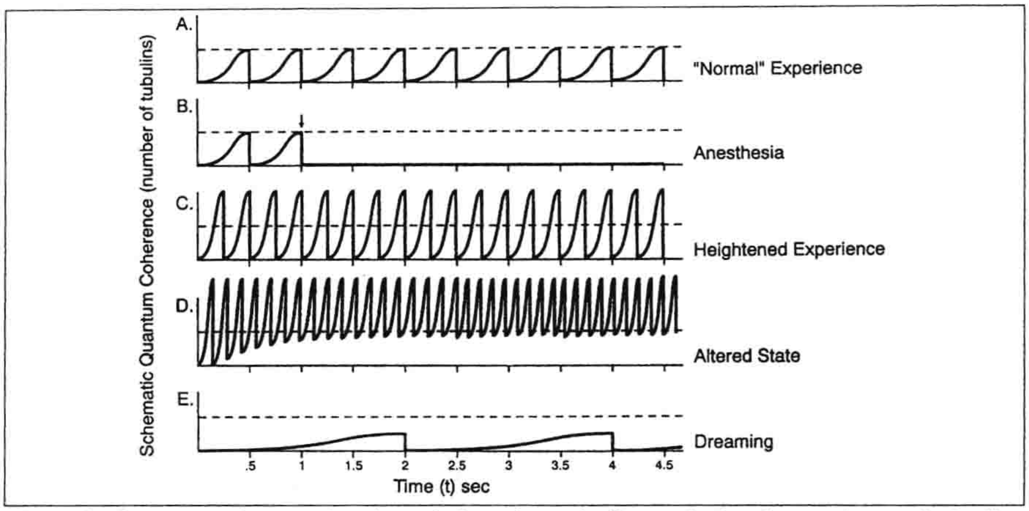


Figure 8. Quantum coherence in microtubules schematically graphed on longer time scale for five different states related to consciousness. Area under each curve equivalent in all cases. A. Normal experience: as in Figure 8. B. Anaesthesia: anaesthetics bind in hydrophobic pockets and prevent quantum delocalizability and coherent superposition (e.g. Louria and Hameroff, 1996). C. Heightened Experience: increased sensory experience input (for example), increases rate of emergence of quantum coherent superposition. **Orch Or** threshold is reached faster (e.g. 250 msec), and **Orch Or** frequency is doubled. D. Altered State: even greater rate of emergence of quantum coherence due to sensory input and other factors promoting quantum state (e.g. meditation, psychedelic drug, etc.). Predisposition to quantum state results in baseline shift and only partial collapse so that conscious experience merges with normally sub–conscious quantum computing mode. E. Dreaming: prolonged quantum coherence time. (Highly schematic.)

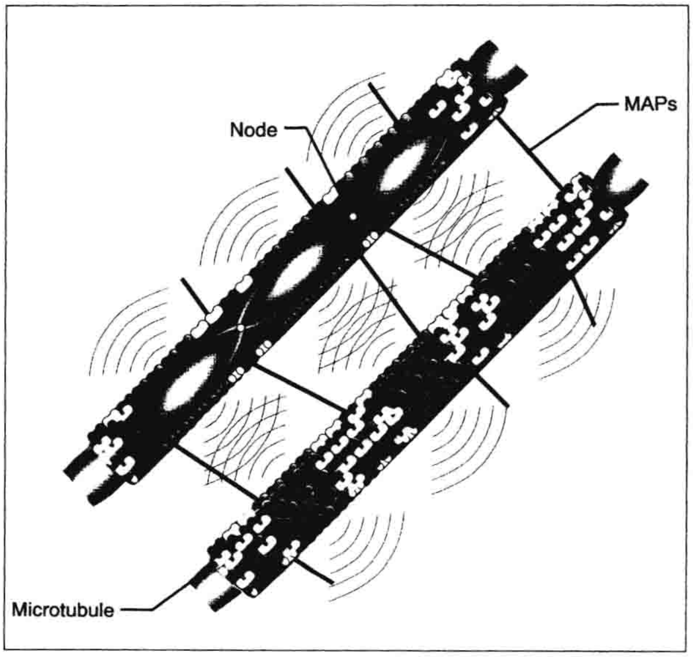


Figure 9. Quantum coherence in microtubules. Having emerged from resonance in classical automaton patterns, quantum coherence non–locally links superpositioned tubulins (grey), within and among microtubules. Upper microtubule: cutaway view shows coherent photons generated by quantum ordering of water on tubulin surfaces, propagating in microtubule waveguide. MAP (microtubule-associated protein), attachments breach isolation and prevent quantum coherence; MAP attachment sites thus act as 'nodes' which tune and orchestrate quantum oscillations and set possibilities and probabilities for collapse outcomes ('orchestrated objective reduction': **Orch OR**).

We envisage that attachments of MAPs on microtubules 'tune' quantum oscillations, and 'orchestrate' possible collapse outcomes (Figure 9). Thus we term the particular *self*-organizing **OR** occurring in MAP-connected microtubules, and relevant to consciousness, orchestrated objective reduction ('**Orch OR**'). **Orch OR** events are thus *self*-selecting processes in fundamental space–time geometry. If experience is truly a component of fundamental space–time, **Orch OR** may indeed begin to to address the 'hard problem' of consciousness.

Summary of the 'Orch OR' Model for Consciousness

The full details of this model are given in Hameroff and Penrose (1996). The picture we are putting forth involves the following ingredients:

(1) Aspects of quantum theory (e.g. quantum coherence) and of the suggested physical phenomenon of quantum wave function '*self*-collapse' (objective reduction: **OR** — Penrose, 1994; 1996) are essential for consciousness, and occur in cytoskeletal microtubules (MTs) and other structures within each of the brain's neurons.

(2) Conformational states of MT subunits (tubulins) are coupled to internal quantum events, and cooperatively interact with other tubulins in both classical and quantum computation (Hameroff *et al.*, 1992; Rasmussen *et al.*, 1990 — Figures 4, 5 and 6).

(3) Quantum coherence occurs among tubulins in MTs, pumped by thermal and biochemical energies (perhaps in the manner proposed by Fröhlich, 1968; 1970; 1975). Evidence for some kind of coherent excitation in proteins has recently been reported by Vos *et al.* (1993).

It is also considered that water at MT surfaces is 'ordered' — dynamically coupled to the protein surface. Water ordering within the hollow MT core (acting something like a quantum waveguide) may result in quantum coherent photons (as suggested by the phenomena of 'super-radiance'and '*self*-induced transparency' — Jibu *et al.*, 1994; 1995). We require that coherence be sustained (protected from environmental interaction) for up to hundreds of milliseconds by isolation (a) within hollow MT cores, (b) within tubulin hydrophobic pockets, (c) by coherently ordered water, (d) sol-gel layering (Hameroff and Penrose, 1996). Feasibility of quantum coherence in the seemingly noisy, chaotic cell environment is supported by the observation that quantum spins from biochemical radical pairs which become separated retain their correlation in cytoplasm (Walleczek, 1995).

(4) During preconscious processing, quantum coherent superposition/computation occurs in MT tubulins and continues until the mass-distribution difference among the separated states of tubulins reaches a threshold related to quantum gravity. Self-collapse (**OR**) then occurs (Figures 6 & 7).

(5) The **OR** *self*-collapse process results in classical 'outcome states' of MT tubulins which then implement neurophysiological functions. According to certain ideas for **OR** (Penrose, 1994), the outcome states are 'non-computable'; that is they cannot be determined algorithmically from the tubulin states at the beginning of the quantum computation.

(6) Possibilities and probabilities for post**OR** tubulin states are influenced by factors including initial tubulin states, and attachments of microtubule-associated proteins (MAPs) acting as 'nodes' which tune and 'orchestrate' the quantum oscillations (Figure 9). We thus term the *self*-tuning **OR** process in microtubules 'orchestrated objective reduction' — **Orch OR**.

(7) According to the arguments for **OR** put forth in Penrose (1994), superposed states each have their own space–time geometries. When the degree of coherent mass–energy difference leads to sufficient separation of space–time geometry, the system must choose

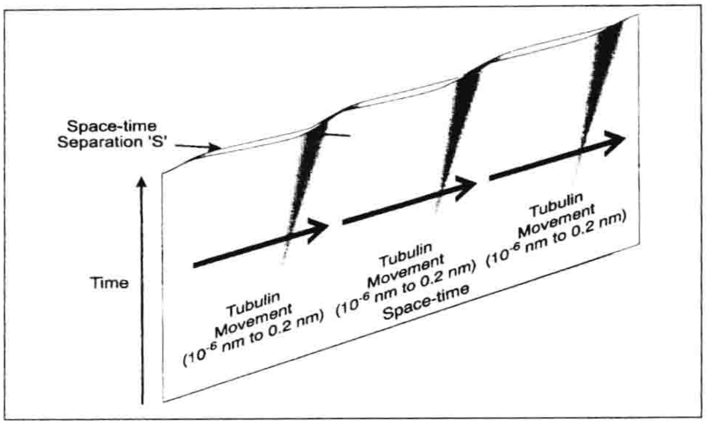

Space-time Separation 'S'
Time
Tubulin Movement (10^{-6} nm to 0.2 nm)
Tubulin Movement (10^{-6} nm to 0.2 nm)
Tubulin Movement (10^{-6} nm to 0.2 nm)
Space-time

Figure 10.

Schematic space–time separation illustration of three superposed tubulins. The space–time differences are very tiny in ordinary terms (10^{-40} nm), but relatively large mass movements (e.g. hundreds of tubulin conformations, each moving from 10^{-6} nm to 0.2 nm), indeed have precisely such very tiny effects on the space–time curvature.

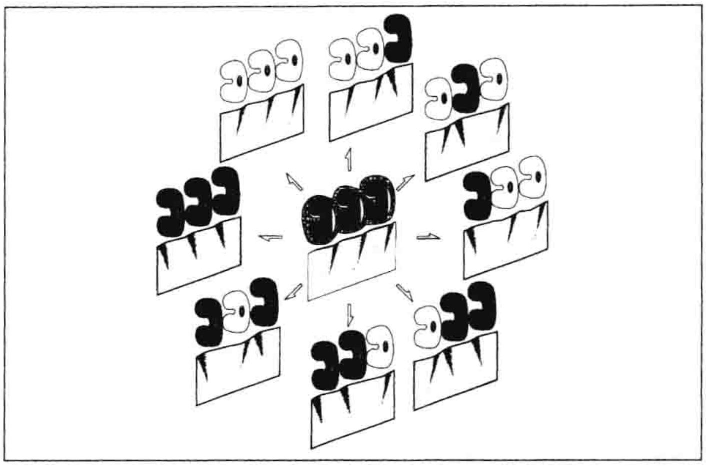


Figure 11.

Centre:
Three superposed tubulins (e.g. Figure 4), with corresponding schematic space–time separation illustrations (Figures 1 and 10). Surrounding the superposed tubulins are the eight possible post-reduction 'eigenstates' for tubulin conformation, and corresponding space–time geometry.

and decay (reduce, collapse) to a single universe state. Thus **Orch OR** involves *self*-selections in fundamental space–time geometry (Figures 10 & 11).

(8) To quantify the **Orch OR** process, in the case of a pair of roughly equally superposed states, each of which has a reasonably well-defined mass distribution, we calculate the gravitational *self*-energy ***E*** of the difference between these two mass distributions, and then obtain the approximate lifetime ***T*** for the superposition to decay into one state or the other by the formula $T = \hbar / E$. Here, ħ is Planck's constant over 2π. We call ***T*** the coherence time for the superposition (how long the coherence is sustained). If we assume a coherence time $T = 500$ msec (shown by Libet *et al.*, 1979, and others to be a relevant time for preconscious processing), we calculate ***E***, and determine the number of MT tubulins whose coherent superposition for 500 msec will elicit **Orch OR**. This turns out to be about 10^9 tubulins.

(9) A typical brain neuron has roughly 10^7 tubulins (Yu and Baas, 1994). If, say, 10 % of tubulins within each neuron are involved in the quantum coherent state, then roughly 10^3 (one thousand) neurons would be required to sustain coherence for 500 msec, at which time the quantum gravity threshold is reached and **Orch OR** then occurs.

(10) We consider each *self*-organized **Orch OR** as a single conscious event; cascades of such events would constitute a 'stream' of consciousness. If we assume some form of excitatory input (e.g. you are threatened, or enchanted) in which quantum coherence emerges faster, then, for example, 10^{10} coherent tubulins could **Orch OR** after 50 msec,

or 10^{11} after 5 msec (e.g. Figure 8c). Turning to see a bengal tiger in your face might perhaps elicit 10^{12} in 0.5 msec, or more tubulins, faster. A slow emergence of coherence (your forgotten phone bill) may require longer times. A single electron would require more than the age of the universe.
(11) Quantum states are non-local (because of quantum entanglement — or 'Einstein-Podolsky-Rosen' (EPR) effects), so that the entire non-localized state reduces all at once. This can happen if the mass movement that induces collapse takes place in a small region encompassed by the state, or if it takes place uniformly over a large region. Thus, each instantaneous **Orch OR** could 'bind' various superpositions which may have evolved in separated spatial distributions and even over different time scales, but whose net displacement *self*-energy reaches threshold at a particular moment. Information is bound into an instantaneous event (a 'conscious now'). Cascades of **Orch OR**s could then represent our familiar 'stream of consciousness', and create a 'forward' flow of time (Aharonov and Vaidman, 1990; Elitzur, 1996; Tollaksen, 1996).

It may be interesting to compare our considerations with subjective viewpoints that have been expressed with regard to the nature of the progression of conscious experience. For example, support for consciousness consisting of sequences of individual, discrete events is found in Buddhism; trained meditators describe distinct 'flickerings' in their experience of reality (Tart, 1995). Buddhist texts portray consciousness as 'momentary collections of mental phenomena', and as 'distinct, unconnected and impermanent moments which perish as soon as they arise'. Each conscious moment successively becomes, exists, and disappears — its existence is instantaneous, with no duration in time, as a point has no length. Our normal perceptions, of course, are seemingly continuous, presumably as we perceive 'movies' as continuous despite their actual makeup being a series of frames. Some Buddhist writings even quantify the frequency of conscious moments. For example the Sarvaastivaadins (von Rospatt, 1995) described 6,480,000 'moments' in 24 hours (an average of one 'moment' per 13.3 msec), while other Buddhist writings decribe one moment per 0.13 msec (Conze, 1988), and some Chinese Buddhism as one 'thought' per 20 msec. These accounts, including variations in frequency, seem to be consistent with our proposed **Orch OR** events. For example a 13.3 msec preconscious interval would correspond with an **Orch OR** involving 4×10^{10} coherent tubulins, a 0.13 msec interval would correspond with 4×10^{12} coherent tubulins, and a 20 msec interval with 2.5×10^{10} coherent tubulins. Thus Buddhist 'moments of experience', Whitehead 'occasions of experience', and our proposed **Orch OR** events seem to correspond tolerably well with one another.

The **Orch OR** model thus appears to accommodate some important features of consciousness:

1) control/regulation of neural action
2) pre-conscious to conscious transition
3) non-computability
4) causality
5) binding of various (time scale and spatial) superpositions into instantaneous 'now'
6) a 'flow' of time, and
7) a connection to fundamental space–time geometry in which experience may be based.

Conclusion: What Is It Like to Be a Worm?

The **Orch OR** model has the implication that an organism able to sustain quantum coherence among, for example, 10^9 tubulins for 500 msec might be capable of having conscious experience. More tubulins coherent for a briefer period, or fewer for a longer period ($\boldsymbol{E} = £/\boldsymbol{T}$) will also have conscious events. Human brains appear capable of, for example, 10^{12} tubulin, 0.5 msec 'bengal tiger experiences', but what about simpler organisms?

From an evolutionary standpoint, introduction of a dynamically functional cytoskeleton (perhaps symbiotically from spirochetes, e.g. Margulis, 1975) greatly enhanced eukaryotic cells by providing cell movement, internal organization, separation of chromosomes and numerous other functions. As cells became more specialized with extensions like axopods and eventually neural processes, increasingly larger cytoskeletal arrays providing transport and motility may have developed quantum coherence via the Fröhlich mechanism as a by-product of their functional coordination.

Another possible scenario for emergence of quantum coherence leading to **Orch OR** and conscious events is 'cellular vision'. Albrecht-Buehler (1992) has observed that single cells utilize their cytoskeletons in 'cellular vision' — detection, orientation and directional response to beams of red/infrared light. Jibu *et al.* (1995) argue that this process requires quantum coherence in microtubules and ordered water, and Hagan (1995) suggests the quantum effects/cellular vision provided an evolutionary advantage for cytoskeletal arrays capable of quantum coherence. For whatever reason quantum coherence emerged, one could then suppose that, one day, an organism achieved sufficient microtubule quantum coherence to elicit **Orch OR**, and had a 'conscious' experience.

At what level of evolutionary development might this primitive consciousness have emerged? A single cell organism like *Paramecium* is extremely clever, and utilizes its cytoskeleton extensively. Could a paramecium be conscious? Assuming a single paramecium contains, like each neuronal cell, 10^7 tubulins, then for a paramecium to elicit **Orch OR**, 100% of its tubulins would need to remain in quantum coherent superposition for nearly a minute. This seems unlikely.

Consider the nematode worm *C elegans*. It's 302 neuron nervous system is completely mapped. Could *C elegans* support **Orch OR**? With 3×10^9 tubulins, *C elegans* would require roughly one third of its tubulins to sustain quantum coherent superposition for 500 msec. This seems unlikely, but not altogether impossible. If not *C elegans*, then perhaps *Aplysia* with a thousand neurons, or some higher organism. **Orch OR** provides a theoretical framework to entertain such possibilities.

Would a primitive **Orch OR** experience be anything like ours? If *C elegans* were able to *self*-collapse, what would it be like to be a worm? (cf. Nagel, 1974). A single, 10^9 tubulin, 500 msec **Orch OR** in *C elegans* should be equal in gravitational *self*-energy terms (and thus perhaps, experiential intensity) to one of our 'everyday experiences'. A major difference is that we would have many **Orch OR** events sequentially (up to, say, 50 per second) whereas *C elegans* could generate, at most, 2 per second. *C elegans* would also presumably lack extensive memory and associations, and have poor sensory data, but nonetheless, by our criteria a 10^9 tubulin, 500 msec **Orch OR** in *C elegans* could be a conscious experience: a mere smudge of known reality, the next space–time move.

Consciousness has an important place in the universe. **Orch OR** in microtubules is a model depicting consciousness as sequences of non-computable *self*-selections in fundamental space time geometry. If experience is a quality of space–time, then **Orch OR** indeed begins to address the 'hard problem' of consciousness in a serious way.

References

Aharonov, Y. and Vaidman, L. (1990), 'Properties of a quantum system during the time interval between two measurements', *Phys. Rev. A*, **41**, p. 11.

Albrecht–Buehler, G. (1992), 'Rudimentary form of ''cellular vision'' ', *Cell. Biol.* **89**, pp. 8288–92.

Beck, F. and Eccles, J.C. (1992), 'Quantum aspects of brain activity and the role of consciousness', *Proc. Natl. Acad. Sci. USA,* **89** (23), pp. 11357–61.

Chalmers, D. (1995), 'Facing up to the problem of consiousness', *Journal of Consciousness Studies*, **2** (3), pp. 200–19 (reprinted in this volume).

Chalmers, D. (1996), *The Conscious Mind (New York:* Oxford University Press).

Conze, E. (1988), *Buddhist Thought in India*, Louis de La Vallee Poussin (trans.), Abhidharmako'sabhaa.syam: English translation by Leo M. Pruden, 4 vols (Berkeley), pp. 85–90.

DeBrabander, M. (1982), 'A model for the microtubule organizing activity of the centrosomes and kinetochores in mammalian cells', *Cell Biol. Intern. Rep.*, **6**, pp. 901–15.

Diósi, L. (1989), 'Models for universal reduction of macroscopic quantum fluctuations', *Phys. Rev. A*, **40**, 1165–74.

Elitzur, A. (1996), 'Time and consciousness: The uneasy bearing of relativity theory on the mind–body problem', in Hameroff *et al.* (1996).

Everett, H. (1957), 'Relative state formulation of quantum mechanics', *Rev. Mod. Physics,* **29**, pp. 454–62. Reprinted in *Quantum Theory and Measurement*, ed. J.A. Wheeler and W.H. Zurek (Princeton: Princeton University Press, 1983).

Fröhlich, H. (1968), 'Long–range coherence and energy storage in biological systems', *Int. J. Quantum Chem.*, **2**, pp. 641–9.

Fröhlich, H. (1970), 'Long range coherence and the actions of enzymes', *Nature,* **228**, p. 1093.

Fröhlich, H. (1975), 'The extraordinary dielectric properties of biological materials and the action of enzymes', *Proc. Natl. Acad. Sci.,* **72**, pp. 4211–15.

Ghirardi, G.C., Grassi, R. and Rimini, A. (1990), 'Continuous–spontaneous reduction model involving gravity', *Phys. Rev. A*, **42**, pp. 1057–64.

Ghirardi, G.C., Rimini, A. and Weber, T. (1986), Unified dynamics for microscopic and macroscopic systems', *Phys. Rev. D*, **34**, p. 470.

Goswami, A. (1993), *The Self-Aware Universe: How Consciousness Creates the Material World* (New York: Tarcher/Putnam).

Hagan, S. (1995), Personal communication.

Hameroff, S.R., Dayhoff, J.E., Lahoz–Beltra, R., Samsonovich, A. and Rasmussen, S. (1992), 'Conformational automata in the cytoskeleton: models for molecular computation' *IEEE Computer* (October Special Issue on Molecular Computing), pp. 30–9.

Hameroff, S.R., Kaszniak, A. and Scott, A.C. (eds. 1996), *Toward a Science of Consciousness – The First Tucson Discussions and Debates* (Cambridge, MA: MIT Press).

Hameroff, S.R. and Penrose, R. (1995), 'Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness', *Neural Network World*, **5** (5), 793–804.

Hameroff, S.R. and Penrose, R. (1996), 'Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness', in Hameroff *et al.* (1996).

Hameroff, S.R. and Watt, R.C. (1982), 'Information processing in microtubules', *J. Theor. Biol.*, **98**, pp. 549–61.

Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H. and Yasue, K. (1994), 'Quantum optical coherence in cytoskeletal microtubules: implications for brain function', *BioSystems*, **32**, pp. 195–209.

Jibu, M., Yasue, K. and Hagan, S. (1995), 'Water laser as cellular ''vision'' ', submitted.

Károlyházy, F., Frenkel, A., and Lukacs, B. (1986), 'On the possible role of gravity on the reduction of the wave function', in *Quantum Concepts in Space and Time,* ed. R. Penrose and C.J. Isham (Oxford: Oxford University Press).

Károlyházy, F. (1966), 'Gravitation and quantum mechanics of macroscopic bodies', *Nuovo Cim.* A, **42**, p. 390.
Károlyházy, F. (1974), 'Gravitation and quantum mechanics of macroscopic bodies', *Magyar Fizikai Polyoirat*, **12**, p. 24.
Kibble, T.W.B.(1981), 'Is a semi–classical theory of gravity viable?' in *'Quantum Gravity 2: A Second Oxford Symposium'*, ed. C.J. Isham, R. Penrose and D.W. Sciama (Oxford: Oxford University Press).
Libet, B., Wright, E.W. Jr., Feinstein, B. and Pearl, D.K. (1979), 'Subjective referral of the timing for a conscious sensory experience', *Brain*, **102**, pp. 193–224.
Louria, D. and Hameroff, S. (1996),' Computer simulation of anesthetic binding in protein hydrophobic pockets', in Hameroff *et al.* (1996).
Marshall, I.N. (1989), 'Consciousness and Bose–Einstein condensates', *New Ideas in Psychology*, **7**, 73–83.
Margulis, L. (1975), *Origin of Eukaryotic Cells* (New Haven: Yale University Press).
Nagel, T. (1974), 'What is it like to be a bat?' *The Philosophical Review*, **83**, pp. 435–50. Reprinted in *The Mind's I. Fantasies and Reflections on Self and Soul,* ed. D.R. Hofstadter and D.C. Dennett (New York: Basic Books, 1981).
Pearle, P. (1989), 'Combining stochastic dynamical state vector reduction with spontaneous localization', *Phys. Rev. D*, **13**, 857–68.
Pearle, P. and Squires, E. (1995), 'Gravity, energy conservation and parameter values in collapse models', *Durham University preprint,* DTP/95/13.
Penrose, R. (1987), 'Newton, quantum theory and reality', in *300 Years of Gravity*, ed. S.W. Hawking and W. Israel (Cambridge: Cambridge University Press).
Penrose, R. (1989), *The Emperor's New Mind* (Oxford: Oxford University Press).
Penrose, R. (1993), 'Gravity and quantum mechanics', in *General Relativity and Gravitation. Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina 28 June–4 July 1992. Part 1: Plenary Lectures*, ed. R.J. Gleiser, C.N. Kozameh and O.M. Moreschi (Bristol: Institute of Physics Publications).
Penrose, R. (1994), *Shadows of the Mind* (Oxford: Oxford University Press).
Penrose, R. and Hameroff, S.R., 'What gaps? Reply to Grush and Churchland', *Journal of Consciousness Studies*, **2** (2), pp. 99–112.
Penrose, R. (1996), 'On gravity's role in quantum state reduction', *Gen.Rel.Grav.*, to appear May 1996.
Percival, I.C. (1995), 'Quantum space–time fluctuations and primary state diffusion', *Proc. Roy. Soc. Lond.* A, **451**, pp. 503–13.
Rasmussen, S., Karampurwala, H., Vaidyanath, R., Jensen, K.S. and Hameroff, S. (1990), 'Computational connectionism within neurons: A model of cytoskeletal automata subserving neural networks', *Physica D*, **42**, pp. 428–49.
Rensch, B. (1960), *Evolution Above the Species Level* (New York: Columbia University Press).
Russell, B. (1954), *The Analysis of Matter* (New York: Dover).
Schrödinger, E (1935), Die gegenwarten situation in der quantenmechanik. *Naturwissenschaften*, **23**, pp. 807–12, 823–8, 844–9. (Translation by J. T. Trimmer, 1980, in *Proc. Amer. Phil. Soc.,* **124**, pp. 323–38.) In *Quantum Theory and Measurement*, ed. J.A. Wheeler and W.H. Zurek (Princeton: Princeton University Press, 1983).
Shimony, A. (1993), *Search for a Naturalistic World View – Volume II. Natural Science and Metaphysics* (Cambridge: Cambridge University Press).
Spinoza, B. (1677), *Ethica in Opera quotque reperta sunt,* 3rd edition, ed. J. van Vloten and J.P.N. Land (Netherlands: Den Haag)
Stubenberg, L. (1996), 'The place of qualia in the world of science', in Hameroff *et al.* (1996).
Tart, C.T. (1995), Personal communication and information gathered from 'Buddha–1 newsnet'.
Tollaksen, J. (1996), 'New insights from quantum theory on time, consciousness, and reality', in Hameroff *et al.* (1996).
Tusz‰yski, J., Hameroff, S.R., Sataric, M.V., Trpisová, B. and Nip, M.L.A. (1995), 'Ferroelectric behavior in microtubule dipole lattices: implications for information processing, signalling and assembly/ disassembly', *J. Theor. Biol.*, **174**, pp. 371–80.
von Rospatt, A. (1995), *The Buddhist Doctrine of Momentariness: A survey of the origins and early phase of this doctrine up to Vasubandhu* (Stuttgart: Franz Steiner Verlag).

Vos, M.H., Rappaport, J., Lambry, J. Ch., Breton, J. and Martin, J.L. (1993), 'Visualization of coherent nuclear motion in a membrane protein by femtosecond laser spectroscopy', *Nature*, **363**, pp. 320–5.

Walleczek, J. (1995), 'Magnetokinetic effects on radical pairs: a possible paradigm for understanding sub–kT magnetic field interactions with biological systems', in *Biological Effects of Environmental Electromagnetic Fields* (Advances in Chemistry, No. 250), ed. M. Blank (Washington, DC: American Chemical Society Books, in press).

Wheeler, J.A. (1957), 'Assessment of Everett's ''relative state'' formulation of quantum theory', *Revs. Mod. Phys.,* **29**, pp. 463–5.

Wheeler, J.A. (1990), 'Information, physics, quantum: The search for links', in *Complexity, Entropy, and the Physics of Information*, ed. W. Zurek (Addison–Wesley).

Whitehead, A.N. (1929), *Science and the Modern World* (New York: Macmillan).

Whitehead, A.N. (1933), *Process and Reality* (New York: Macmillan).

Yu, W. and Baas, P.W. (1994), 'Changes in microtubule number and length during axon differentiation', *J. Neuroscience*, **14** (5), pp. 2818–29.

247

ORCHESTRATED REDUCTION OF QUANTUM COHERENCE IN BRAIN MICROTUBULES: A MODEL FOR CONSCIOUSNESS?

Toward a science of consciousness: The first Tucson discussions and debates. Eds S. R. Hameroff, A. W. Kaszniak, and A. C. Scott, 507–40. MIT Press, Cambridge, 1996. With S. Hameroff. Also published in *Mathematics and Computers in Simulation* **40** (1996) 453–80

See Chapter 246.

Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness?
Toward a science of consciousness: The first Tucson discussions and debates: Eds S. R. Hameroff, A. W. Kaszniak, and A. C. Scott, 507–40. 1996. With S. Hameroff. Also published in *Mathematics and Computers in Simulation* **40** (1996) 453–80

ELSEVIER

Mathematics and Computers in Simulation 40 (1996) 453–480

MATHEMATICS
AND COMPUTERS
IN SIMULATION

Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness

Stuart Hameroff [a,*], Roger Penrose [b]

[a] *Departments of Anesthesiology and Psychology, University of Arizona, 1501 North Campbell Avenue, Tucson, Arizona 85724, USA*
[b] *Rouse Ball Professor of Mathematics, University of Oxford, Oxford, United Kingdom*

Abstract

Features of consciousness difficult to understand in terms of conventional neuroscience have evoked application of quantum theory, which describes the fundamental behavior of matter and energy. In this paper we propose that aspects of quantum theory (e.g. quantum coherence) and of a newly proposed physical phenomenon of quantum wave function "self-collapse" (objective reduction: **OR** – Penrose, 1994) are essential for consciousness, and occur in cytoskeletal microtubules and other structures within each of the brain's neurons. The particular characteristics of microtubules suitable for quantum effects include their crystal-like lattice structure, hollow inner core, organization of cell function and capacity for information processing. We envisage that conformational states of microtubule subunits (tubulins) are coupled to internal quantum events, and cooperatively interact (compute) with other tubulins. We further assume that macroscopic coherent superposition of quantum-coupled tubulin conformational states occurs throughout significant brain volumes and provides the global binding essential to consciousness. We equate the emergence of the microtubule quantum coherence with pre-conscious processing which grows (for up to 500 ms) until the mass–energy difference among the separated states of tubulins reaches a threshold related to quantum gravity. According to the arguments for **OR** put forth in Penrose (1994), superpositioned states each have their own space–time geometries. When the degree of coherent mass–energy difference leads to sufficient separation of space–time geometry, the system must choose and decay (reduce, collapse) to a single universe state. In this way, a transient superposition of slightly differing space–time geometries persists until an abrupt quantum → classical reduction occurs. Unlike the random, "subjective reduction" (**SR**, or **R**) of standard quantum theory caused by observation or environmental entanglement, the **OR** we propose in microtubules is a *self*-collapse and it results in particular patterns of microtubule-tubulin conformational states that regulate neuronal activities including synaptic functions. Possibilities and probabilities for post-reduction tubulin states are influenced by factors including attachments of microtubule-associated proteins (MAPs) acting as "nodes" which tune and "orchestrate" the quantum oscillations. We thus term the self-tuning **OR** process in microtubules "orchestrated objective reduction" ("**Orch OR**"), and calculate an estimate for the number of tubulins (and neurons) whose coherence for relevant time periods (e.g. 500 ms) will elicit **Orch OR**. In providing a connection among (1) pre-conscious to conscious transition, (2) fundamental space–time notions, (3) non-computability, and (4) binding of various (time scale and spatial) reductions into an instantaneous event ("conscious now"), we believe **Orch OR** in brain microtubules is the most specific and plausible model for consciousness yet proposed.

---

* Corresponding author. Tel.: (520) 626-2116. Fax: (520) 626-2689. E-Mail: srh@ccit.arizona.edu.

Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness?

Toward a science of consciousness: The first Tucson discussions and debates: Eds S. R. Hameroff, A. W. Kaszniak, and A. C. Scott, 507–40. 1996. With S. Hameroff. Also published in *Mathematics and Computers in Simulation* **40** (1996) 453–80

CHAPTER 247

Keywords: Consciousness; Cytoskeleton; Microtubules; Quantum gravity; Quantum theory

1. Introduction

Current neurophysiological explanations of consciousness suggest that it is a manifestation of emergent firing patterns of neuronal groups involved in either specific networks (e.g. [33, 34, 53, 54]), coherent 40–80 Hz firing (e.g. [19, 45, 92]) and/or attentional scanning circuits (e.g. [7, 8, 18, 30]). But even precise correlation of neuronal firing patterns with cognitive activities fails to address perplexing differences between mind and brain including the "hard problem" of the nature of our inner experience (e.g. Chalmers, this Volume).[1] In this paper we apply certain aspects of quantum theory (quantum coherence) and a new physical phenomenon described in Penrose [79] of wave function self-collapse (objective reduction: **OR**) to specific, essential structures within each neuron: cytoskeletal microtubules. Table 1 summarizes how quantum coherence and **OR** occurring in microtubules **(Orch OR)** can potentially address some of the inexplicable features of consciousness.

Quantum theory describes the surprising behavior at a fundamental level of matter and energy which comprise our universe. At the base of quantum theory is the wave/particle duality of atoms and their components. As long as a quantum system such as an atom or sub-atomic particle remains

Table 1
Aspects of consciousness difficult to explain by conventional neuroscience and possible quantum solutions

Problematic feature of consciousness	Possible quantum solutions
Unitary sense: "binding problem"	(1) Non-local quantum coherence; Indivisible macroscopic quantum state (e.g. Bose–Einstein condensate); (2) Instantaneous *self*-collapse of superpositioned states (**Orch OR**).
Transition from pre-conscious/sub-conscious to conscious processes	(1) Sub- and pre-conscious occur in quantum computing mode; (2) Automatic, autonomic functions occur in classical computing mode; (3) Quantum → classical transition. (Wave function "self" collapse – **Orch OR** – is intrinsic to consciousness).
Non-computable, non-algorithmic logic	**Orch OR** is non-computable.
(Apparent) non-deterministic "free will"	Non-computable, but non-random wave function self-collapse (**Orch OR**).
Essential nature of human experience	(1) Wave function self collapse (**Orch OR**) from incompatible superposition of separated space–times; (2) Pre-conscious → conscious transition; (3) Effectively instantaneous "now" (**Orch OR**) collapse.

---

[1] Citations to "This Volume" refer to *Toward a Science of Consciousnes*: The First Tucson Discussions and Debates, (1996) S. Hameroff, A. Kaszniak, A. Scott (eds), MIT Press, Cambridge.

isolated from its environment, it behaves as a "wave of possibilities" and exists in coherent "superposition" (with complex number coefficients) of many possible states.

There are differing views as to how quantum superposed states (wave functions) are "collapsed" or "reduced" to a single, classical state (Table 2). The conventional quantum theory view (Copenhagen interpretation) is that the quantum state reduces by environmental entanglement, measurement or conscious observation (subjective reduction: **SR**, or **R**). Precisely where a quantum particle is and how it is moving when observed is "indeterminate" and, according to the Copenhagen interpretation, results in random measured values. We take the view [79] that, to address this issue, a new physical ingredient (objective reduction: **OR**) is needed in which coherent quantum systems can "self-collapse" by growing and persisting to reach a critical mass/time/energy threshold related to quantum gravity. In the **OR** scheme, the collapse outcomes ("eigenstates") need not be random, but can reflect (in some non-computable way) a quantum computation occurring in the coherent superposition state.

Another feature of quantum systems is quantum inseparability, or non-locality, which implies that all quantum objects that have once interacted are in some sense still connected! When two quantum systems have interacted, their wave functions become "phase entangled" so that when one system's wave function is collapsed, the other system's wave function, no matter how far away, instantly collapses as well. The non-local connection ("quantum entanglement") is instantaneous, independent of distance and implies that the quantum entities, by sharing a wave function, are indivisible.

Where and how in the brain can quantum effects occur? Warm, wet and noisy, the brain at first glance seems a hostile environment for delicate quantum phenomena which generally demand isolation and cold stillness (superconductors), or energy pumping of crystals (lasers). Nonetheless, various authors have implicated ion channels, ions themselves, DNA, pre-synaptic grids and cytoskeletal microtubules as somehow mediating "standard" quantum effects. In a dualist context, Beck and Eccles [9] proposed that an external "conscious self" might influence the apparently random quantum effects acting on neurotransmitter release at the pre-synaptic grid within each neural axon. Stapp [88] has suggested that (**SR**) wave function collapse in neurons is closely related to consciousness in the brain. In our view, cytoskeletal microtubules are the most likely sites for quantum coherence, **OR** and consciousness.

Table 2
Descriptions of wave function collapse

Context	Cause of collapse (Reduction)	Description	Acronym
Quantum coherent superposition	No collapse	Evolution of the wave function (Schrödinger equation)	U
Conventional quantum theory (Copenhagen interpretation)	Environmental entanglement, Measurement, Conscious observation	Reduction; Subjective reduction	**R, SR**
New physics (Penrose, 1994)	Self-collapse – quantum gravity induced (Penrose, Diosi, etc.)	Objective reduction	**OR**
Consciousness (present paper)	Self-collapse, quantum gravity threshold in microtubules orchestrated by MAPs, etc.	Orchestrated objective reduction	**Orch OR**

Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness?

Toward a science of consciousness: The first Tucson discussions and debates: Eds S. R. Hameroff,
A. W. Kaszniak, and A. C. Scott, 507–40. 1996. With S. Hameroff. Also published in
Mathematics and Computers in Simulation **40** (1996) 453–80

Networks of self-assembling protein polymers, the cytoskeleton within neurons establishes neuronal form, maintains synaptic connections, and performs other essential tasks (Fig. 1). The major cytoskeletal components are microtubules, hollow cylindrical polymers of individual proteins known as tubulin. Microtubules are interconnected by linking proteins (microtubule-associated proteins: MAPs) to other microtubules and cell structures to form cytoskeletal lattice networks (Fig. 2).

Traditionally viewed as the cell's "bone-like" scaffolding, microtubules and other cytoskeletal structures also appear to fill communicative and information processing roles. Theoretical models suggest how conformational states of tubulins within microtubule lattices can interact with neighboring tubulins to represent, propagate and process information as in molecular-level "cellular automata" computing systems (e.g. [50, 52, 82]).

In this paper, we present a model linking microtubules to consciousness using quantum theory as viewed in a particular "realistic" way, as described in *Shadows of the Mind* [79]. In our model, quantum coherence emerges, and is isolated, in brain microtubules until the differences in mass–energy distribution among superpositioned tubulin states reaches a threshold related to quantum

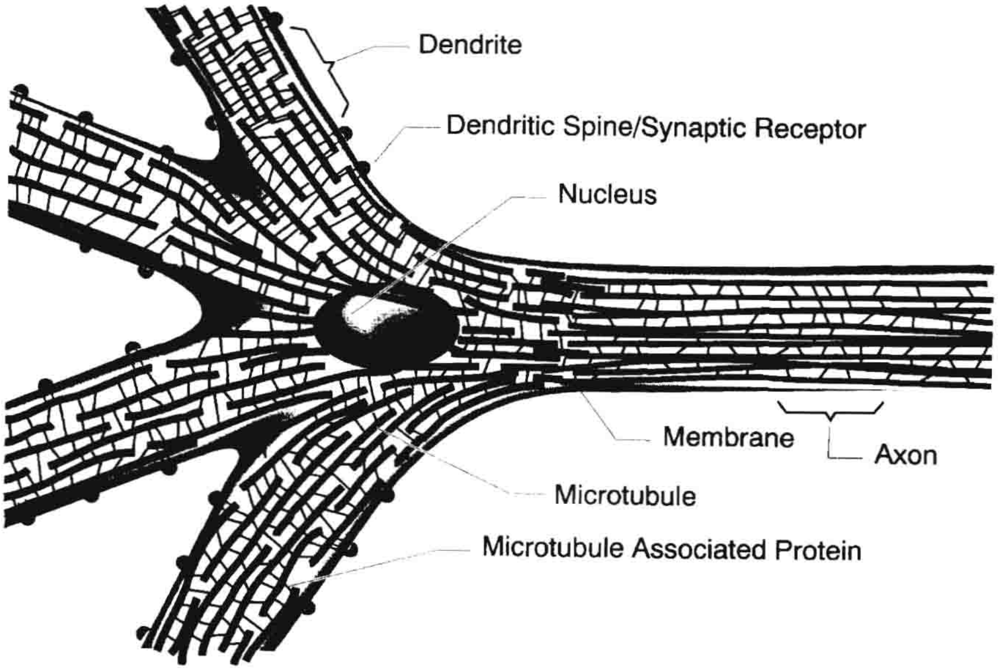

Dendrite
Dendritic Spine/Synaptic Receptor
Nucleus
Membrane
Axon
Microtubule
Microtubule Associated Protein

Fig. 1. Schematic of central region of neuron (distal axon and dendrites not shown) showing parallel arrayed microtubules interconnected by MAPs. Microtubules in axons are lengthy and continuous, whereas in dendrites they are interrupted and of mixed polarity. Linking proteins connect microtubules to membrane proteins including receptors on dendritic spines.

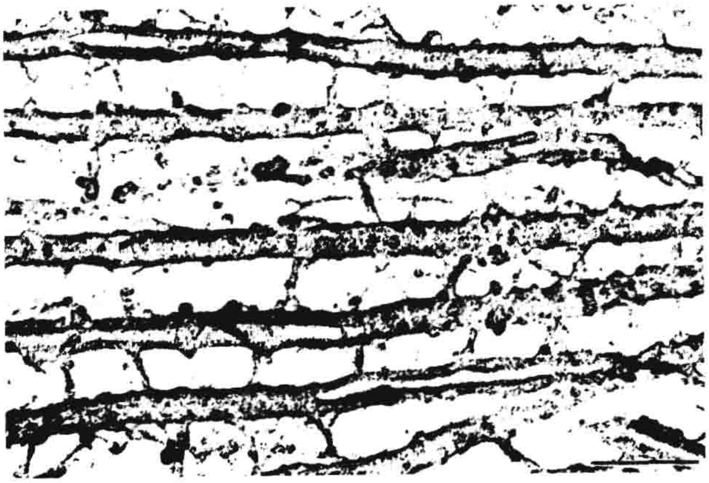


Fig. 2. Immunofluorescent micrograph of neuronal microtubules interconnected by MAPs. Scale bar: 100 nm. (With permission from Hirokawa, 1991).

gravity. The resultant self-collapse (**OR**), irreversible in time, creates an instantaneous "now" event. Sequences of such events create a flow of time, and consciousness.

We envisage that attachments of MAPs on microtubules "tune" quantum oscillations, and "orchestrate" possible collapse outcomes. Thus we term the particular **OR** occurring in MAP-connected microtubules, and relevant to consciousness, as "orchestrated objective reduction" (**Orch OR**).

2. Collapse of the wave function

The boundary between the microscopic, quantum world and the macroscopic, classical world remains enigmatic. Behavior of wave-like, quantum-level objects can be satisfactorily described in terms of a deterministic, unitarily evolving process (e.g. state vector evolving according to the Schrödinger equation) denoted by **U**. Large-scale (classical) systems seem to obey (different) computable deterministic laws. The transition when system effects are magnified from the small, quantum scale to the large, classical scale (measurement process) chooses a particular "eigenstate" (one state of many possible states). According to the conventional Copenhagen interpretation of quantum theory, the "choice" of eigenstate is purely random. The non-computable **R** process is known in various contexts as collapse of the wave function, quantum jump, Heisenberg event and/or state reduction.

Von Neumann, Schrödinger and others in the 1930s supposed that quantum collapse, or **R** effectively occurred when a quantum system interacted with its environment, was otherwise "measured" or consciously observed. Exactly why and how collapse occurs, and how eigenstates are determined, are unknown and indicate a gap in physics knowledge: **R** is not taken to be an objectively real, independent phenomenon in the standard Copenhagen interpretation.

A number of physicists have argued in support of specific models (or of general schemes) in which the rules of standard **U**-quantum mechanics are modified by the inclusion of some additional procedure according to which **R** does become an objectively real process. The relevant procedure of any such specific scheme is here denoted by **OR** (objective reduction). In *Shadows of the Mind*, Penrose [79] describes **OR** in which quantum coherence grows until it reaches a critical threshold related to quantum gravity, and then abruptly self-collapses. Other schemes for **OR** include those due to dynamics [43, 75], and those which are based on gravitational effects, such as [27, 42, 58, 78]. Recent work [76] lends some considerable support, on general and observational grounds, for a gravitational **OR** scheme. There are also strong arguments from other directions [77, 78] supporting a belief that the appropriate union of general relativity with quantum mechanics will lead to a significant change in the latter theory (as well as in the former – which is generally accepted). There is also some tentative, but direct, evidence in favor of this union being a non-computable theory (e.g. [41]; Deutsch, unpublished; cf. [79]). **OR** in microtubules relevant to consciousness was first considered in somewhat general terms in *Shadows of the Mind*. Here we shall adopt a fairly specific proposal for **OR** (in accordance with [27, 42, 79]) applied quantitatively in microtubules in which emergence of quantum coherence U and subsequent **OR** are "guided" and "tuned" ("orchestrated") by connecting MAPs. We thus elaborate a model of "orchestrated **OR**" (**Orch OR**) in microtubules which may support consciousness.

An important feature of **OR** (and **Orch OR**) is that non-computable aspects arise only when the quantum system becomes large enough that its state undergoes self-collapse, rather than its state collapsing because its growth forces entanglement with its environment. Because of the random nature of environment, the **OR** action resulting from growth-induced entanglement would be indistinguishable from the random **SR**, or **R** process of standard quantum theory.

Consciousness, it is argued, requires non-computability [78, 79]. In standard quantum theory there is no non-computable activity, the **R** process being totally random. The only readily available apparent source of non-computability is **OR** (and **Orch OR**) self-collapse. An essential feature of consciousness might then be a large-scale quantum-coherent state maintained for a considerable time. **OR** (**Orch OR**) then takes place because of a sufficient mass displacement in this state, so that it indulges in a self-collapse which somehow influences or controls brain function. Microtubules seem to provide easily the most promising place for these requirements.

3. Microtubules and the cytoskeleton

Ideal properties for quantum brain structures relevant to consciousness might include: (1) high prevalence, (2) functional importance (for example regulating neural connectivity and synaptic function), (3) periodic, crystal-like structure with long-range order, (4) ability to be transiently isolated from external interaction/observation, (5) functionally coupled to quantum-level events, and (6) suitable for information processing. Membranes, membrane proteins, synapses, DNA and other candidates have some, but not all, of these characteristics. Cytoskeletal microtubules do appear to have the requisite properties.

Interiors of living eukaryotic cells (including the brain's neurons and glia) are organized by integrated networks of protein polymers called the cytoskeleton (e.g. [28, 49]). In addition to "bonelike" support, these dynamic self-organizing networks appear also to play roles as each cell's circulatory and nervous systems.

The cytoskeleton consists of microtubules (MTs), actin filaments, intermediate (neuro) filaments and microtubule-associated proteins (MAPs) which, among other duties, link these parallel structures into networks (Figs. 1 and 2).

The most prominent cytoskeletal component, MTs are hollow cylinders 25 nanometers (nm = 10^{-9} m) in diameter whose lengths vary and may be quite long within some nerve axons. MT cylinder walls are comprised of 13 longitudinal protofilaments which are each a series of subunit proteins known as tubulin (Fig. 3). Each tubulin subunit is a polar, 8 nm dimer which consists of two slightly different classes of 4 nm, 55 000 dalton monomers known as α and β tubulin. The tubulin dimer subunits within MTs are arranged in a hexagonal lattice which is slightly twisted, resulting in differing neighbor relationships among each subunit and its six nearest neighbors, and helical pathways which repeat every 3, 5 and 8 rows.

MTs, as well as their individual tubulins, have dipoles with negative charges localized toward α monomers [22]. Thus MTs are "electrets": oriented assemblies of dipoles which are predicted to have piezoelectric [6, 69] and ferroelectric [90] properties. Biochemical energy is provided to the cytoskeleton in at least two ways: tubulin bound GTP is hydrolyzed to GDP in MTs, and MAPs are phosphorylated. Each tubulin has a large hydrophobic region [3], a non-polar pocket of amino acid side groups which interact by van der Waals forces and can support quantum level electron delocalizability (e.g. Louria and Hameroff, this Volume).

MTs self-assemble and disassemble (e.g. [61]). The different configurations they assume by their assembly and MAP attachments determine cell form and function including synaptic connections in neurons. Cell architecture (and synaptic connections) can quickly adapt by MT disassembly, and subsequent reassembly and MAP network formation in another shape or direction. Many organized

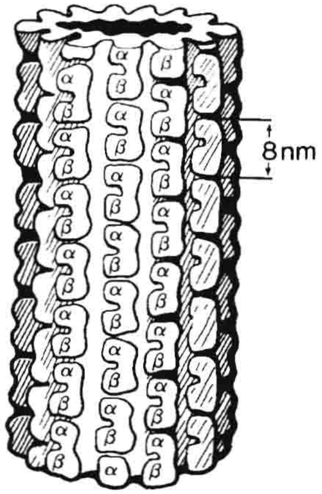


Fig. 3. Microtubule structure from x-ray crystallography [2]. Tubulin subunits are 8 nm dimers comprised of alpha and beta monomers.

cytoskeletal functions are carried out by MAPs. Some MAPs (dynein, kinesin) act as motors and carry material along microtubules (axoplasmic transport).

Several types of studies suggest cytoskeletal involvement in cognition. For example long term potentiation (LTP) is a form of synaptic plasticity that serves as a model for learning and memory in mammalian hippocampal cortex. LTP requires MAP-2, a dendrite-specific, MT-crosslinking MAP which is dephosphorylated as a result of synaptic membrane receptor activation (e.g. [48]). In cat visual cortex, MAP-2 is dephosphorylated when visual stimulation occurs [4]. Auditory Pavlovian conditioning elevates temporal cortex MAP-2 activity in rats [95]. Phosphorylation/dephosphorylation of MAP-2 accounts for a large proportion of brain biochemical energy consumption (e.g. [89]) and is involved in functions which include strengthening specific networks, such as potentiating excitatory synaptic pathways in rat hippocampus [73]. The mechanism for regulating synaptic function appears related to rearrangement of MAP-2 connections on MTs [12, 35].

Other types of evidence also link the cytoskeleton with cognitive function. Production of tubulin and MT activities correlate with peak learning, memory and experience in baby chick brains [72]. When baby rats first open their eyes, neurons in their visual cortex begin producing vast quantities of tubulin [20]. In animals whose brains are temporarily deprived of oxygen, the degree of cognitive damage correlates with decrease in measured levels of dendritic MAP-2 [62]. Bensimon and Chernat [11] showed that selective damage of MTs in animal brains by the drug colchicine causes defects in learning and memory which mimic the symptoms of Alzheimer's disease (in which neuronal cytoskeleton entangles). Matsuyama and Jarvik [70] have linked Alzheimer's disease to microtubule dysfunction, and Mandelkow and Mandelkow [67] and others have pinpointed the axonal MAP "tau protein" as the tangle-causing defect.

How might the cytoskeleton signal and process information? Tubulin can undergo several types of conformational changes (e.g. [16, 31]). Roth and Pihlaja [83] suggested that patterns of tubulin conformation within MTs represented information. In one example of tubulin conformational change observed in single protofilament chains, one monomer can shift 27 degrees from the dimer's vertical axis [71]. Whether such mechanical deformation occurs in tubulin within intact MTs is unknown; neighbor tubulins in the MT lattice might be expected to constrain movement. However, cooperativity among tubulins bound loosely in the MT lattice by hydrophobic forces could coordinate conformational changes, and support propagation of wave-like signals in MTs. Vassilev [91] demonstrated signal transmission along tubulin chains formed between excitable membranes. A number of models of signaling and information processing within MTs and other cytoskeletal components have been suggested. These include propagating tubulin conformational changes [5], ion transfer [14], sequential phosphorylation/dephosphorylation along MT tubulins [81], tensegrity [94], non-linear soliton waves along MTs [1, 5, 84] and "cellular automaton" behavior due to electrostatic dipole coupling among tubulin lattice neighbors (e.g. [82]).

4. Microtubule information processing

4.1. Protein conformation

Proteins have conformational transitions at many time and size scales [59]. For example small side chains move in the picosecond to femtosecond time scale (10^{-12}–10^{-15}), but conformational

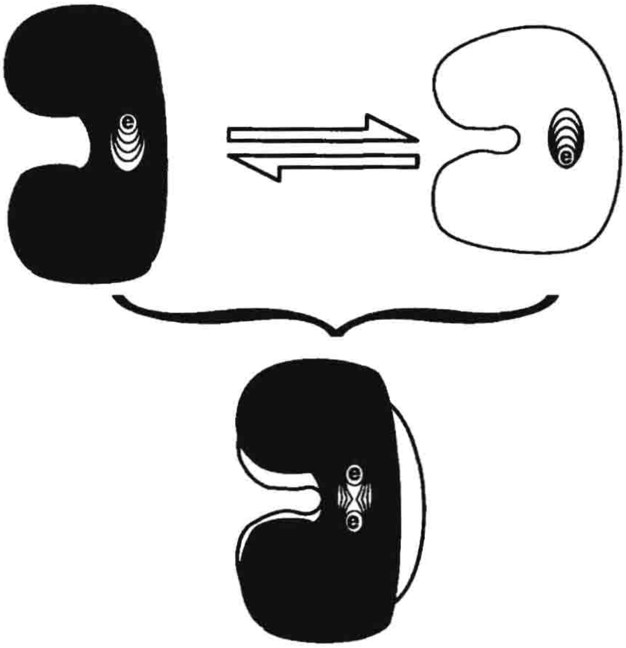


Fig. 4. Schematic model of tubulin states. Top: Two states of microtubule subunit protein "tubulin" in which quantum event (electron localization) within a hydrophobic pocket is coupled to protein conformation. Bottom: Tubulin in quantum coherent superposition of both states.

transitions in which proteins move globally and upon which protein function generally depends occur in the nanosecond (10^{-9} s) to 10 picosecond (10^{-11} s) time scale. Related to cooperative movements of smaller regions, hydrogen bond rearrangements and charge redistributions such as dipole oscillations, these global changes linked to protein function (signal transduction, ion channel opening, enzyme action, etc.) may be regulated by a variety of factors including phosphorylation, ATP or GTP hydrolysis, ion fluxes, electric fields, ligand binding, and "allosteric" influences by neighboring protein conformational changes. Noting the extraordinary dielectric strength of proteins (their ability to sustain a voltage), Fröhlich [36–38] proposed that the various factors determining protein conformation were integrated through a quantum level dipole oscillation within each protein's hydrophobic region (Fig. 4).

4.2. *Fröhlich's coherent pumped phonons*

In addition to linking protein conformation to hydrophobic quantum events, Herbert Fröhlich, an early contributor to the understanding of superconductivity, also predicted quantum coherence in living cells (based on earlier work by Penrose and Onsager [80]). Fröhlich theorized that sets of protein dipoles in a common electromagnetic field (e.g. proteins within a polarized membrane, subunits within an electret polymer like microtubules) undergo coherent conformational excitations if energy is supplied. Fröhlich postulated that biochemical and thermal energy from the surrounding "heat bath" provides such energy. Cooperative, organized processes leading to coherent excitations emerged, according to Fröhlich, because of structural coherence of hydrophobic dipoles in a common voltage gradient.

Coherent excitation frequencies on the order of 10^9–10^{11} Hz (identical to the time domain for functional protein conformational changes, and in the microwave or gigaHz spectral region) were deduced by Fröhlich who termed them acousto-conformational transitions, or coherent (pumped) phonons. Such coherent states are termed Bose–Einstein condensates in quantum physics and have been suggested by Marshall ([68]; this Volume) to provide macroscopic quantum states which support the unitary binding of consciousness.

Experimental evidence for Fröhlich-like coherent excitations in biological systems includes observation of gigaHz-range phonons in proteins [39], sharp-resonant non-thermal effects of microwave irradiation on living cells [47], gigaHz induced activation of microtubule pinocytosis in rat brain [74] and Raman spectroscopy detection of Fröhlich frequency energy [40].

4.3. Cellular automata in microtubules

Computational systems in which complex signaling and patterns emerge from local activities of simple subunits are called cellular automata. Their essential features are: (1) At a given time, each subunit is in one of a finite number of states (usually two for simplicity); (2) The subunits are organized according to a fixed geometry; (3) Each subunit communicates only with neighboring subunits; the size and shape of the neighborhood are the same for all cells; (4) A universal "clock" provides coherence such that each subunit may change to a new state at each "clock tick"; (5) Transition rules for changing state depend on each subunit's "present" state and those of its neighbors. Depending on initial conditions (starting patterns), simple neighbor transition rules can lead to complex, dynamic patterns capable of computation. Von Neumann [93] proved mathematically that cellular automata could function as Turing machines.

In a series of simulations (e.g. [51, 82]) Fröhlich's excitations were used as a clocking mechanism and electrostatic dipole coupling forces as "transition rules" for cellular automata behavior by dynamic conformational states of tubulins within MTs. As in the top of Fig. 4, the two monomers of each tubulin dimer are considered to share a mobile electron within a hydrophobic pocket which is oriented either more toward the α-monomer (alpha state) or more toward the β-monomer (beta state) with associated changes in tubulin conformation at each "Fröhlich coherent" time step (e.g. 10^{-9}–10^{-11} s). The net electrostatic force (f_{net}) from the six surrounding neighbors acting on each tubulin can then be calculated as

$$f_{\text{net}} = \frac{e^2}{4\pi\varepsilon} \sum_{i=1}^{6} \frac{y_i}{r_i^3},$$

where y_i and r_i are inter-tubulin distances, e is the electron charge, and ε is the average protein permittivity. MT automata simulations (Fig. 5) show conformational pattern behaviors including standing waves, oscillators and gliders traveling one dimer length (8 nm) per time step (10^{-9}–10^{-11} s) for a velocity range of 8–800 meters per seconds (m/s), consistent with the velocity of propagating nerve action potentials. MT automata patterns can thus represent and process

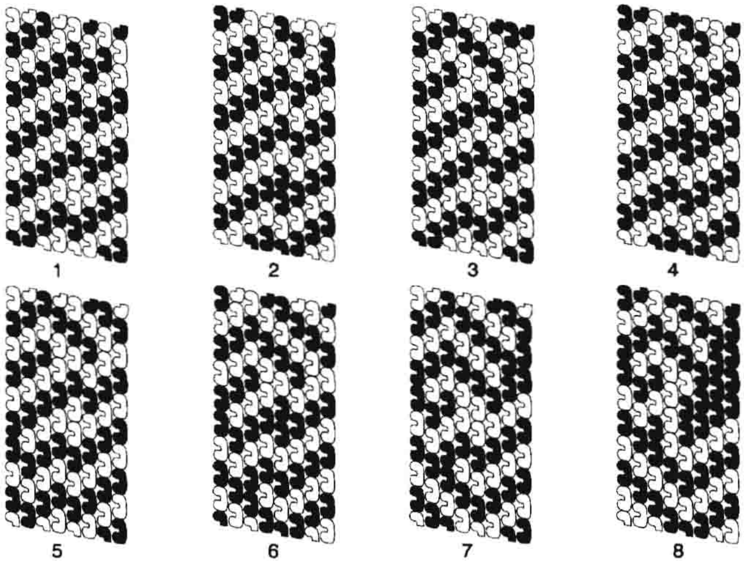


Fig. 5. Microtubule automaton simulation [82]. Black and white tubulins correspond to states shown in Fig. 4. Eight nanosecond time steps of a segment of one microtubule are shown in "classical computing" mode in which patterns move, evolve, interact and lead to emergence of new patterns.

information through each cell; gliders may convey signals which regulate synaptic strengths, represent binding sites for MAPs (and thus neuronal and synaptic connectionist architecture) or material to be transported. Information could become "hardened" in MTs by tubulin modifications or stored in neurofilaments via MAPs.

MT conformational automata patterns provide a further level of computational complexity within each of the brain's neurons. However, by considering only classical computing and local neighbor interactions, microtubule automata fail to address the problematic features of consciousness for which quantum theory holds promise.

5. Wave function self-collapse in microtubules: OR and Orch OR

5.1. The collapse criterion and conscious thought

Consider a quantum superposition $w|A\rangle + z|B\rangle$ (where w and z are complex numbers) of two macroscopically distinguishable quantum states $|A\rangle$ and $|B\rangle$. In standard quantum theory and in the absence of environmental entanglement, this superposition would persist forever. If, after a time t, $|A\rangle$ would have evolved to $|A_t\rangle$ and $|B\rangle$ would have evolved to $|B_t\rangle$, then $w|A\rangle + z|B\rangle$ must evolve to $w|A_t\rangle + z|B_t\rangle$. (This is a feature of the linear nature of **U**.)

According to the present **OR** criterion, such macroscopic superpositions are regarded as *unstable* even without environmental entanglement. Therefore the state $w|A\rangle + z|B\rangle$ will decay in a certain time scale T, to either $|A\rangle$ or $|B\rangle$, with relative probabilities $|w|^2:|z|^2$. (This is analogous to the situation with an *unstable radioactive particle*, with a lifetime T and two separate decay modes $|A\rangle$ or $|B\rangle$, whose branching ratios are $|w|^2:|z|^2$.) The idea is that the states $|A\rangle$ and $|B\rangle$ each correspond to clearly defined energy distributions (and to well defined space–time geometries), whereas the combination $w|A\rangle + z|B\rangle$ does not (and so would lead to superpositions of different space–time geometries – a particularly awkward situation from the physical point of view!). According to a number of authors (e.g. [27, 42, 43, 75; Penrose, 1993; 79]), the gravitational self-energy E of the difference between the mass distributions involved in $|A\rangle$ and in $|B\rangle$ will determine spontaneous reduction (at time T) of the superposed combination $w|A\rangle + z|B\rangle$ to either $|A\rangle$ or $|B\rangle$.

We view $|A\rangle$ and $|B\rangle$ as representing two conformationally coupled quantum states of each tubulin within microtubules, and $w|A\rangle + z|B\rangle$ as the superposition of those states (e.g. Fig. 4). We envisage that time T, at which the state $w|A\rangle + z|B\rangle$ for each tubulin will decay to either $|A\rangle$ or $|B\rangle$ should relate to the transition between pre-conscious and conscious events.

We assume that pre- and sub-conscious processing corresponds with quantum coherent superposition which can perform "quantum computing" [78]. A number of authors (e.g. [10, 25, 26, 32]) have proposed that quantum coherence can implement multiple computations simultaneously, in parallel, according to quantum linear superposition: the quantum state then "collapses" to a particular result. A state which "self-collapses" (**OR**) will have an element of non-computability, even though evolution of its quantum coherence had been linear and computable. A quantum superposed state collapsed by external environment or observation (**SR**, or **R**) lacks a non-computable element, and would thus be unsuitable for consciousness. Large scale quantum coherence occurring among tubulins (e.g. via electrons in hydrophobic pockets arrayed in the microtubule lattice, or ordered water within hollow MT cores) could take on aspects of a quantum computer in pre-conscious and sub-conscious modes.

We also assume that "non-conscious" autonomic processes correspond with classical, non-quantum computing by microtubule conformational automata. Thus an **OR** transition from quantum, pre-conscious processing, to classical, non-conscious processing may be closely identified with consciousness itself.

But what is consciousness? According to the principles of **OR** [79], superpositioned states each have their own space–time geometries. When the degree of coherent mass-energy difference leads to sufficient separation of space–time geometry, the system must choose and decay (reduce, collapse) to a single universe state, thus preventing "multiple universes" (e.g. Wheeler, 1957). In this way, a transient superposition of slightly differing space–time geometries persists until an abrupt quantum → classical reduction occurs and one or the other is chosen. Thus consciousness may involve self-perturbations of space–time geometry.

The extent of space–time superposition causing self-collapse is related to quantum gravity, and equal to one in "absolute units".

Absolute units convert all physical measurement into pure, dimensionless numbers (cf. [79, pp. 337–339]). This is done by choosing units of length, mass and time so that the following constants take the value of unity:

c (speed of light) $= 1$,

$\hbar$ (Planck's constant divided by 2π) $= 1$,

G (gravitational constant) $= 1$.

Physical quantities relevant to our calculations in absolute units:

$$\text{second} = 1.9 \times 10^{43},$$

$$\text{nanometer} = 6.3 \times 10^{25},$$

$$\text{mass of nucleon ("dalton")} = 7.8 \times 10^{-20},$$

$$\text{fermi (strong interaction size, diameter of nucleon)} = 6.3 \times 10^{19}.$$

Using absolute units, we can ask how many tubulins in quantum coherent superposition for how long will self-collapse (**Orch OR**)?

The gravitational self-energy E for a quantum superposition of mass whose displacement for a given time T sufficiently perturbs space–time for **OR** (**Orch OR**) is taken from the "uncertainty principle":

$$E = \frac{\hbar}{T},$$

where $\hbar$ is Planck's constant divided by 2π, and T is the coherence time.

We estimate T the coherence time from research by Libet [64] and others (e.g. [23, 46, 65]) who found the time scale characteristic of pre-conscious to conscious transitions to be about 500 ms. This is similar to other estimates in the 100–200 ms range (e.g. Koch, this Volume). Hence $T = 500\,\text{ms}$ (half second) seems appropriate from experimental results for at least some functionally significant transitions.

In absolute units,

$$T = 500\,\text{ms} = (0.5)\,1.9 \times 10^{43} \cong 10^{43}.$$

Since $\hbar = 1$ in absolute units,

$$E = T^{-1} = 10^{-43}.$$

This is the gravitational self-energy E for which n_t tubulins displaced in quantum coherent superposition for 500 ms will self-collapse (**OR, Orch OR**). To determine n_t, we calculate the gravitational displacement self-energy E_t for one tubulin. We assume the tubulin conformational movement displaces its mass by a distance r which is 1/10 the 2 nm radius of the tubulin monomer, or 0.2 nm. In absolute units

$$r = 0.2(6.3 \times 10^{25}) \cong 10^{25}.$$

As illustrated in Fig. 6, the distribution of mass m for each tubulin conformational change may be considered as either: (1) two protein spheres, (2) two granular arrays of atoms, (3) two granular arrays of nucleons (protons and neutrons).

5.1.1. Protein spheres

Consider the tubulin dimer as two uniform (monomer) spheres (Fig. 6(a)). As a reasonable approximation, the energy E_t for one tubulin dimer is twice the displacement energy of one (monomer) sphere position in the gravitational field of its other position (twice because there are two

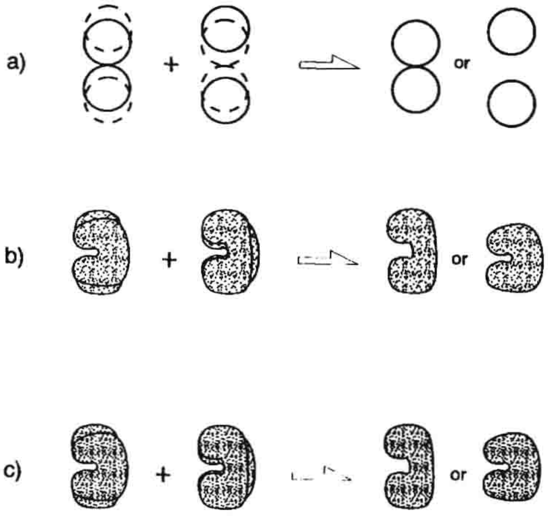


Fig. 6. Three treatments of mass distribution of tubulin. (a) 2 protein spheres, (b) 2 granular arrays of (carbon) atoms, (c) 2 granular arrays of nucleons.

spheres per tubulin). The mass m_t of the tubulin monomer sphere is 55 000 daltons (nucleons). In absolute units

$$m_t = 5.5 \times 10^4 (7.8 \times 10^{-20}) \cong 4 \times 10^{-15}.$$

Since the displacement is less than the sphere radius, we need a detailed calculation to obtain the gravitational self-energy E_m of the difference between the displaced mass distributions for each tubulin monomer, considered as a uniform sphere. Taking the sphere to have a radius a and the distance of displacement to be r, we find, as the result of a double integration

$$E_m = m^2 G\left(\frac{r^2}{2a^3} - \frac{3r^3}{16a^4} + \frac{r^5}{160a^6}\right),$$

where G is Newton's universal constant of gravitation equal to one in absolute units. Here we have $m = \frac{1}{2} m_t \cong 2 \times 10^{-15}$ and $a \cong 10^{26}$, and we are taking $r = 10^{25}$. We can ignore the higher-order terms, so we obtain, in absolute units

$$E_m \cong \frac{m^2 r^2}{2a^3} \cong \frac{(4 \times 10^{-15} \times 10^{25})^2}{2 \times 10^{78}} = 8 \times 10^{-58}.$$

Thus, approximately, $E_t = 1.6 \times 10^{-57}$ for the gravitational self-energy of displacement for a single dimer. For n_t in absolute units (approximately):

$$n_t = \frac{T^{-1}}{E_t} = \frac{10^{-43}}{1.6 \times 10^{-57}} \cong 6 \times 10^{13} \text{ tubulins.}$$

Considering the tubulin mass distributions simply as protein spheres, we would thus estimate 6×10^{13} tubulins displaced in quantum coherent superposition for 500 ms will self-collapse (**Orch OR**).

5.1.2. Granular arrays of atomic nuclei

Consider tubulin as two arrays of (carbon) atoms (Fig. 6(b)). The mass m_c of one carbon atom (12 nucleons) is its nuclear mass in absolute units:

$$m_c = 12(7.8 \times 10^{-20}) = 10^{-18}.$$

Because the carbon nucleus displacement is greater than its radius (the spheres separate completely), the gravitational self-energy E_c is given by

$$E_c \cong G\frac{m^2}{a_c},$$

where a_c is the carbon nucleus sphere radius equal to 2.5 fermi distances. In absolute units

$$a_c = 2.5(6.3 \times 10^{19}) = 1.6 \times 10^{20}, \qquad E_c = G\frac{m^2}{a_c} = \frac{(10^{-18})^2}{1.6 \times 10^{20}} \cong 10^{-56}.$$

This is the gravitational self-energy in absolute units for displacement of one carbon atom nucleus (12 nucleons). To determine how many carbon nuclei n_t and tubulins n_t displaced for 500 msec will elicit **Orch OR**:

$$E = n_c E_c = T^{-1}, \qquad n_c = \frac{T^{-1}}{E_c} = \frac{10^{-43}}{10^{-56}} = 10^{13} \text{ carbon atoms} = 12 \times 10^{13} \text{ nucleons.}$$

As one tubulin is 110 000 nucleons,

$$n_t = \frac{12 \times 10^{13}}{1.1 \times 10^5} \cong 10^9 \text{ tubulins.}$$

Considering the tubulin mass distribution as atomic (carbon) nuclei, we thus estimate that 10^9 tubulins displaced in coherent superposition for 500 msec will self-collapse, and elicit **Orch OR**.

5.1.3. Granular arrays of nucleons

Consider tubulin as two arrays of nucleons (Fig. 6(c)). The mass m_n of one nucleon in absolute units

$$m_n = 7.8 \times 10^{-20}.$$

Since the nucleon displacement is again greater than its radius (complete separation)

$$E_{\mathrm{n}} = G\frac{m^2}{a_{\mathrm{n}}},$$

where a_{n} is the nucleon radius, or 0.5 fermi

$$a_{\mathrm{n}} = 5(6.3 \times 10^{19}) \cong 3 \times 10^{19}, \qquad E_{\mathrm{n}} = G\frac{m^2}{a_{\mathrm{n}}} = \frac{(7.8 \times 10^{-20})^2}{3 \times 10^{19}} \cong 2 \times 10^{-58}.$$

This is the gravitational self-energy for displacement of one nucleon for 500 ms. ($G = 1$ in absolute units.) To find n_{n}, the number of nucleons whose displacement for 500 ms will elicit **Orch OR**,

$$E = n_{\mathrm{n}} E_{\mathrm{n}} = T^{-1}, \qquad n_{\mathrm{n}} = \frac{T^{-1}}{E_{\mathrm{n}}} = \frac{10^{-43}}{2 \times 10^{-58}} = 5 \times 10^{14}.$$

As one tubulin is 1 10 000 nucleons,

$$n_{\mathrm{n}} = \frac{5 \times 10^{14}}{1.1 \times 10^{5}} \cong 5 \times 10^{9} \text{ tubulins.}$$

Considering tubulin mass as arrays of nucleons, we thus estimate that 5×10^9 tubulins displaced in quantum coherent superposition for 500 msec will self-collapse, **Orch OR**.

Using the three types of tubulin mass distributions (protein spheres, atoms, nucleons), we obtain 6×10^{13}, 10^9 and 5×10^9, respectively for the required number n_t of quantum coherent tubulins displaced for 500 msec to elicit **Orch OR**. Although in approximation all three mass distributions contribute to **Orch OR**, that which gives the highest energy (fewest tubulins, shortest reduction time if T were not fixed) predominates. Thus, 10^9 tubulins is perhaps the best estimate. Section 5.3 considers the possible significance of this number of tubulins.

As discussed in other papers in this Volume (e.g. Elitzur, Tollaksen), consciousness may be linked to creation of an instantaneous "now", and the flow of time. As **Orch OR** is instantaneous, non-computable and irreversible, it can provide "now" moments, and directionality in time. Sequential cascades of **Orch OR** events would then constitute the familiar "stream of consciousness".

The **Orch OR** process in MTs selects patterns (eigenstates of mass distribution) of tubulin conformational states. Fig. 7 illustrates eight possible "eigenstates of mass distribution" for **Orch OR** occurring in three adjacent tubulins. The selected patterns can influence neural function by determining MAP attachment sites and setting initial conditions for MT "cellular automata" information processing. These MT activities can then govern intra-neuronal architecture and synaptic function by modulating sensitivity of membrane receptors (in, for example, dendritic spines), ion channels and synaptic vesicle release mechanisms, communication with genetic material, and regulating axoplasmic transport which accounts for delivery of synaptic material components.

Microtubule associated proteins (MAPs) attached to certain microtubule tubulin subunits would seem likely to communicate the quantum state to the outside "noisy" random environment, and thereby entangle and collapse it (**SR** or **R**, rather than **OR** or **Orch OR**). We therefore presume that

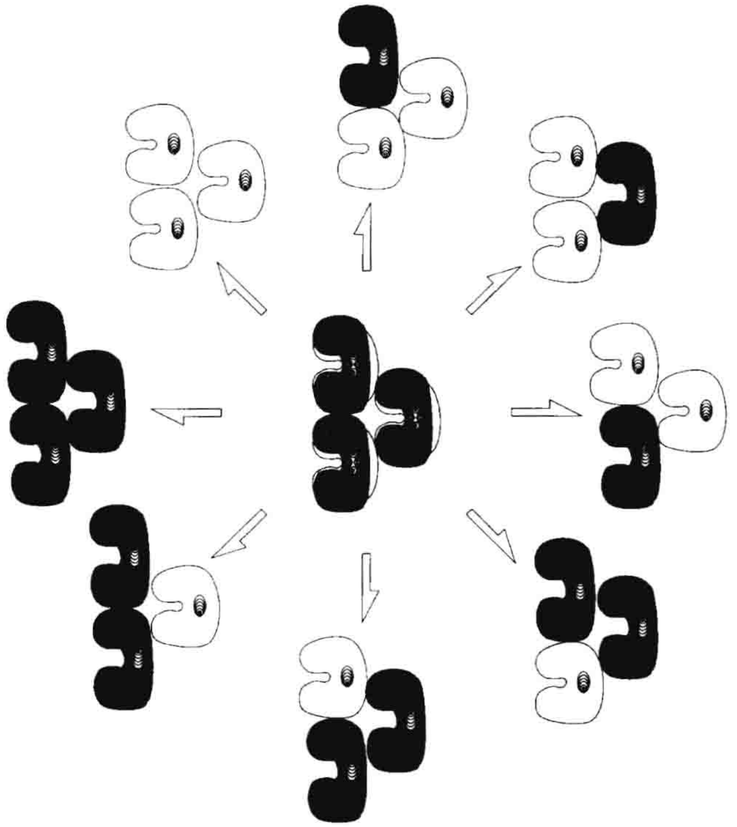


Fig. 7. Possible outcome "eigenstates of mass distribution" for wave function collapse (**Orch OR**) in three tubulins (center) in quantum coherent superposition.

these MAP connections are placed along each MT at sites which are (temporarily at least) inactive with regard to quantum-coupled conformational changes. We envisage that these connection points are, in effect, "nodes" for MT quantum oscillations, and thus "orchestrate" the possibilities and probabilities for MT quantum coherence and subsequent **Orch OR** (Fig. 8). MAP connection points can be regularly placed on MT lattices in super-helical patterns [13, 60] which seems appropriate for their proposed roles as nodes. However in neural MTs, the MAP connection points appear more randomly placed. This would not prevent their acting as "nodes" because the "quantum cellular automaton" activity that we envisage could be extremely complicated, and could well appear to be random.

In addition to MAPs, genetic tubulin variability (e.g. [63]), and "learned, experiential" (post-translational) tubulin modifications (e.g. [44]) can orchestrate **OR**. Accordingly, we term the particular objective reduction (**OR**) occurring in MTs, self-tuned by MAPs and other factors, and relevant to consciousness as "orchestrated objective reduction" (**Orch OR**).

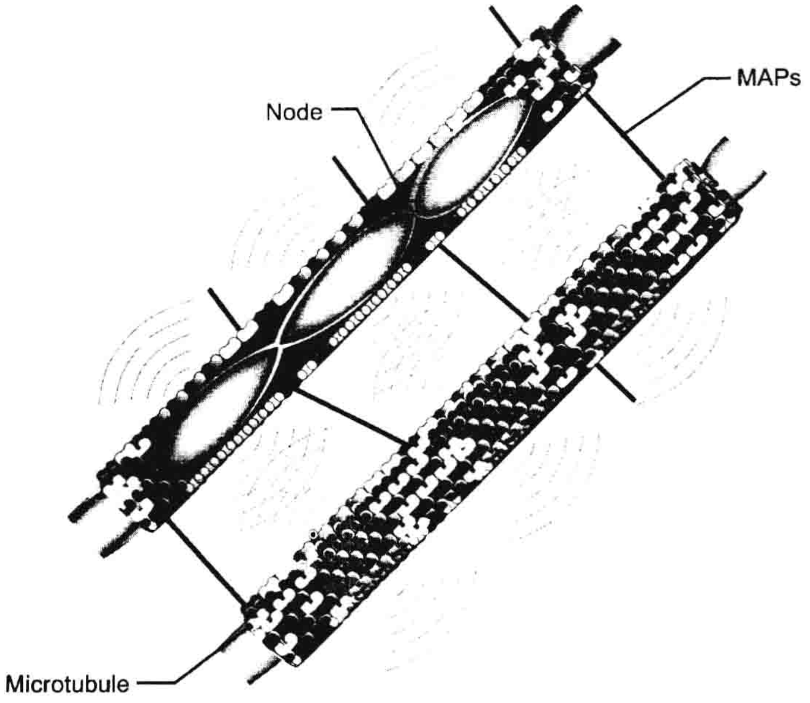


Fig. 8. Quantum coherence in microtubules. Having emerged from resonance in classical automaton patterns, quantum coherence non-locally links superpositioned tubulins (gray) within and among microtubules. Upper microtubule: cutaway view shows coherent photons generated by quantum ordering of water on tubulin surfaces, propagating in microtubule waveguide. MAP (microtubule associated protein) attachments breach isolation and prevent quantum coherence; MAP attachment sites thus act as "nodes" which tune and orchestrate quantum oscillations and set possibilities and probabilities for collapse outcomes (orchestrated objective reduction: **Orch OR**).

Because **OR** phenomena are fundamentally non-local, the coherent superposition phase may exhibit puzzling bidirectional time flow prior to self-collapse (e.g. [1, 78, 79]). As we equate the pre-collapse quantum computing superposition phase to pre-conscious processing, bidirectional time flow could explain the puzzling "backwards time referral" aspects of pre-conscious processing observed by Libet et al. [65].

5.2. How can quantum coherence in microtubules be isolated from environmental entanglement?

When the quantum system under consideration becomes entangled with another system, we must consider the entire state involved. For example $|A\rangle$ might be accompanied by the environment state $|P\rangle$, and $|B\rangle$ with the environment state $|Q\rangle$. Then, in place of the state $w|A\rangle + z|B\rangle$, we have

$$w|A\rangle|P\rangle + z|B\rangle|Q\rangle.$$

We must consider mass movement in $|P\rangle$ and $|Q\rangle$ as well as in $|A\rangle$ and $|B\rangle$, and whenever this dominates, we get, in effect, the random **SR**, or **R** process of conventional quantum measurement theory, rather than the non-computable aspects of **OR** (or **Orch OR**) that would be important for consciousness.

At first glance, the interiors of living cells would seem unlikely sites for quantum effects. "Noisy" thermal motions of cell water ($|P\rangle$) would seem to decohere any quantum coherence and cause random **SR**, or **R** collapse. However, several factors could serve to isolate microtubules and sustain quantum coherence.

5.2.1. Ordered water

Water on cytoskeletal surfaces can be highly "ordered", extending up to 9 layers (about 3 nm) around each microtubule ([17]; Watterson, this Volume). Thermal interactions which would cause decoherence involve energy coupling to oscillations in short range interactions (e.g. hydrogen bonds) among water molecules and groups of water molecules with energy of about kT (10^{-12} s). Interactions of MT surface ordered water molecules with a Fröhlich coherence in MTs are predicted to have a frequency (10^{-14} s) much higher than thermalization energy. Thus MTs may be embedded in "cages" of structured (coherent) water which can act to isolate MT quantum coherence [56].

If a quantized electric field generated by Fröhlich pumped phonons in MTs is comparable to the coherent strength of water, the field penetrates/propagates by "piercing" it (self-focusing, filamentary propagation analogous to Meissner effect in superconductivity). Del Giudice et al. [24] showed this self-focussing should result in filamentous energy beams of radius 15 nm, precisely the inner diameter of microtubules!

5.2.2. Isolation inside microtubule hollow core

Using quantum field theory, Jibu et al. ([56]; this Volume) have modeled the ordering of water molecules and the quantized electromagnetic field confined inside hollow microtubule cores. They predict a specific collective dynamics called super-radiance in which each microtubule can transform incoherent, disordered energy (molecular, thermal, or electromagnetic) into coherent photons within its hollow core. The time for super-radiant photon generation is much shorter than the time needed for the environment to act to disorder the coherence thermally.

5.2.3. Sol–gel states

Cytoplasm is comprised of two phases: sol (solution) and gel (gelatinous). Calcium ions binding to actin and other cytoskeletal polymers convert sol to gel transiently and reversibly (e.g. cytomatrix: [85]. MTs and other MT-associated proteins (i.e. calmodulin) bind/release calcium. Gelatinous phases near adjacent to microtubules could isolate them from thermal effects during quantum coherence.

5.3. The collapse fraction

Here we get a very rough idea of the fraction of brain required for **Orch OR**. In Section 5.1 we obtained a very rough estimate that 10^9 tubulins in quantum coherent superposition for 500 msec are sufficient for **Orch OR**. Yu and Baas [96] measured about 10^7 tubulins per neuron. Thus about 100 neurons whose tubulins were *totally* coherent for 500 msec may be the minimal number for **Orch OR**, and for consciousness. It may be more likely that only a fraction of tubulins within a given neuron becomes coherent. (Global macroscopic states such as superconductivity can result from quantum coherence among only very small fractions of components.) For example if 1% of tubulins within a given set of neurons were coherent for 500 msec, then 10 000 such neurons would be required to elicit **Orch OR**. Thus for 500 msec we get a range for minimal consciousness from hundreds to thousands of neurons. Nervous systems of organisms such as the nematode *C. elegans* contain several hundred neurons. Functional groups of neurons in human congnition are thought to contain thousands. Hebb's [53] "cell assemblies", Eccles' [29] "modules," and Crick and Koch's [19] "coherent set of neurons" are each estimated to contain some 10 000 neurons which may be widely distributed throughout the brain [86].

Rather than always being 500 msec, time T and the number of tubulins n_t may vary, and result in different types of conscious experience. A very intense, sudden input may recruit emergence of quantum coherent tubulins faster so that **Orch OR** occurs sooner ("heightened experience, Fig. 11(c)). For example, 10^{10} coherent tubulins would elicit **Orch OR** in 50 msec, and so on. Lower intensity input patterns develop coherence more slowly, and **Orch OR** occurs later. An instantaneous **Orch OR** may then "bind" various coherent tubulin superpositions whose net displacement energy is T^{-1}, but which may have evolved in separated spatial distributions and over different time

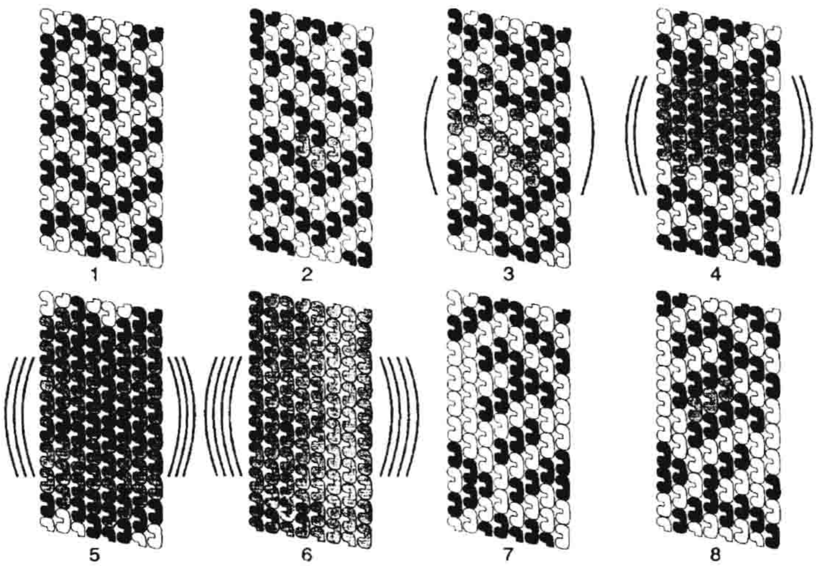


Fig. 9. Microtubule automaton sequence simulation in which classical computing (step 1) leads to emergence of quantum coherent superposition (steps 2–6) in certain (gray) tubulins due to pattern resonance. Step 6 (in coherence with other microtubule tubulins) meets critical threshold related to quantum gravity for self-collapse (**Orch OR**). Consciousness (**Orch OR**) occurs in the step 6 to 7 transition. Step 7 represents the eigenstate of mass distribution of the collapse which evolves by classical computing automata to regulate neural function. Quantum coherence begins to re-emerge in step 8.

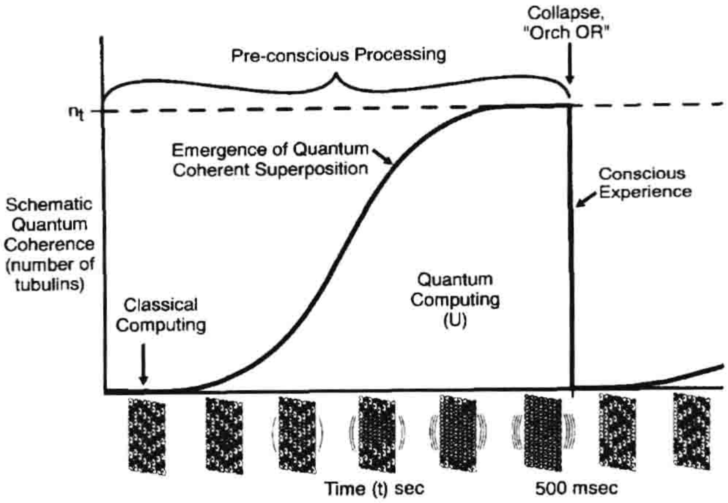


Fig. 10. Schematic graph of proposed quantum coherence (number of tubulins) emerging versus time in microtubules. 500 milliseconds is time for pre-conscious processing (e.g. [65]). Area under curve connects mass–energy difference with collapse time in accordance with gravitational **OR**. This degree of coherent superpositioning of differing space–time geometries leads to abrupt quantum → classical reduction ("self-collapse" or "orchestrated objective reduction: **Orch OR**").

scales into an instantaneous event (a "conscious now"). Cascades of **Orch ORs** can then represent our familiar "stream of consciousness".

6. Summary: orchestrated objective reduction (Orch OR) of quantum coherence in brain microtubules

The picture we are putting forth involves the following ingredients:

(A) A macroscopic state of quantum coherent superposition can exist among tubulin subunits in microtubules across a large proportion of the brain. Plausible candidates for such states include models proposed by:
  (1) Marshall [68] in which Fröhlich pumped phonons induce a Bose–Einstein condensate among proteins distributed throughout the brain.
  (2) Jibu et al. [56] in which ordered water within microtubule hollow cores is coupled to Fröhlich excitations of tubulins in microtubule walls resulting in coherent photons ("super-radiance").
  (3) Kaivarainen [57] in which quantum coherent "flickering" clusters of ordered water within hollow microtubule cores are coupled to (non-Fröhlich) tubulin conformational dynamics and generate coherent photons.
  (4) Conrad (e.g. this Volume) in which conformational states and functional capabilities of proteins are controlled by quantum superposition, dynamics of electrons and hydrogen bonds within them.

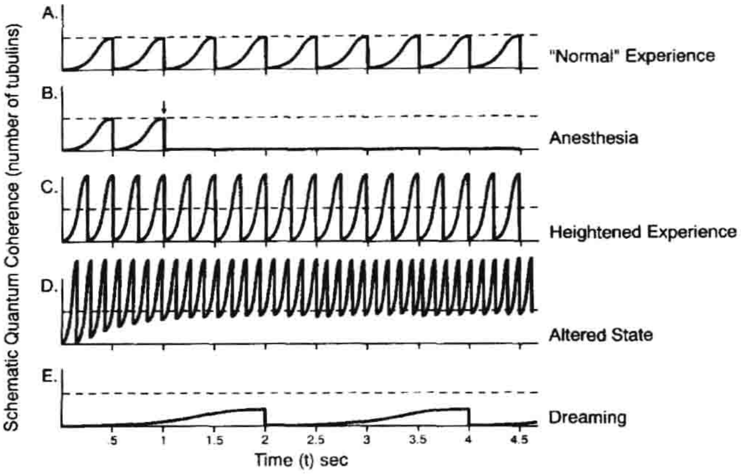


Fig. 11. Quantum coherence in microtubules schematically graphed on longer time scale for 5 different states related to consciousness. Area under each curve equivalent in all cases. A. Normal experience: as in Fig. 8. B. Anesthesia: anesthetics bind in hydrophobic pockets and prevent quantum delocalizability and coherent superposition (see Louria and Hameroff, this Volume). C. Heightened experience: increased sensory experience input (for example) increases rate of emergence of quantum coherent superposition. **Orch OR** threshold is reached faster (e.g. 250 ms) and **Orch OR** frequency is doubled. D. Altered State: even greater rate of emergence of quantum coherence due to sensory input and other factors promoting quantum state (e.g. mediation, psychedelic drug – see Louria and Hameroff, this Volume). Predisposition to quantum state results in baseline shift and only partial collapse so that conscious experience merges with normally sub-conscious quantum computing mode. E. Dreaming: prolonged quantum coherence time.

(B) The quantum coherent state is weakly coupled to conformational activity taking place in tubulins within microtubules. The link occurs by changes in individual electric dipole moments within tubulins. For example movements of a single electron within a hydrophobic pocket centrally placed within each tubulin dimer may couple to the conformational state of the tubulin (Fig. 4).

(C) This combined quantum state among many tubulins is able to maintain itself without significant entanglement with its environment for a relevant period of time (up to 500 msec). We envisage several possible mechanisms which could serve to isolate the MT quantum state from its environment within the neuron. These include (see Section 5.2):
  (1) Shielding by ordered water on tubulin surfaces.
  (2) Isolation within hollow MT inner cores.
  (3) Shielding by gelatinous cytoplasmic layer.

(D) Cooperative interactions among neighboring tubulins in microtubules can signal and process information by computational mechanisms such as cellular automata behavior. We propose two types of microtubule computation:
  (1) Classical computing: conformational patterns propagate through the cytoskeleton to regulate synapses and perform other neural functions (Fig. 5). This mode correlates with non-conscious and autonomic activities.
  (2) Quantum computing: large scale quantum coherence occurs among tubulins (e.g. via electrons in hydrophobic pockets arrayed in the microtubule lattice) and takes on aspects of

a quantum computer (e.g. [10, 25, 26, 32]) where multiple "computations" are performed simultaneously, in parallel, according to quantum linear superposition. We equate quantum computing with pre- and sub-conscious processing.

(E) In the quantum computing mode, changes in dimer conformations involve the movement of mass. According to the arguments and criterion described in Section 5.1, we estimate the time scale T, and calculate the number n_t (of coherent superpositioned tubulins) required for self collapse (objective orchestrated reduction: **Orch OR**) to occur. As we equate quantum computing with pre-conscious processing, we approximate the time scale T to be equivalent in some cases to that found by Libet et al. [65] and others (e.g. [23, 46]) to be characteristic of the transition from pre-conscious to conscious processing (up to 500 ms). For $T = 500$ ms, we get a rough estimate of 10^9 tubulins required for **Orch OR**.

(F) Microtubule associated proteins (MAPs) attached to certain microtubule tubulin subunits would seem likely to communicate the quantum state to the outside "noisy" random environment, and thereby entangle and collapse it (**SR**, or **R**). We therefore presume that these MAP connections are placed along each MT at sites which are (temporarily at least) inactive with regard to quantum-coupled conformational changes. We envisage that these connection points are, in effect, "nodes" for MT quantum oscillations, and (along with genetic and other tubulin modifications) thus "orchestrate" MT quantum coherence and subsequent **OR** (Fig. 8). Accordingly, we term the particular objective reduction (**OR**) occurring in MTs and relevant to consciousness as "orchestrated objective reduction" (**Orch OR**).

(G) The **Orch OR** process selects a new set of tubulin conformational states ("eigenstates of mass distribution") within MTs which can implement and regulate neural function by determining MAP attachment sites and setting initial conditions for "cellular automata" information processing by "classical" conformational transitions. These MT activities can then govern intra-neuronal architecture and synaptic function through modulating sensitivity of membrane receptors, ion channels and synaptic vesicle release mechanisms, communication with genetic material, and regulating axoplasmic transport which accounts for delivery of synaptic material components.

(H) How is a particular conformational pattern within each microtubule chosen in the **Orch OR** process? The Copenhagen quantum interpretation would suggest the selection of states upon (**SR**, or **R**) collapse is purely random. Effects of MAPs, genetic and other tubulin modifications can set the possibilities and probabilities ("orchestrate"), but reduction within that context is non-computable. As described in *Shadows of the Mind* [79], it remains possible that presently unrecognized **OR** (or **Orch OR**) quantum-mathematical logic acting on these programming influences provides a hidden order.

(I) Because **OR** (and **Orch OR**) phenomena are fundamentally non-local, the coherent superposition phase may exhibit puzzling bidirectional time flow prior to self-collapse (e.g. [1, 78, 79]). We equate the pre-collapse quantum computing superposition phase to pre-conscious processing. This could explain the puzzling "backwards time referral" aspects of pre-conscious processing observed by Libet et al. [65]. (Also see Tollaksen, this Volume).

(J) The persistence and global nature of consciousness is seen as a feature of large scale quantum coherent activity taking place across much of the brain. Varieties of **Orch OR** with differing coherence times and amounts of coherent tubulin may blend into conscious thought. Very intense, sudden inputs may recruit emergence of quantum coherent tubulins faster than

500 msec so that **Orch OR** occurs sooner ("heightened experience," Fig. 11(C)). Lower intensity, unexciting input patterns develop coherence more slowly and **Orch OR** occurs later. An instantaneous **Orch OR** may then "bind" disparate tubulin superpositions which may have evolved in separate spatial distributions and over different time scales into an instantaneous conscious "now". Cascades of **Orch OR**s can then represent our familiar "stream of consciousness".

7. Conclusion

Approaches to understanding consciousness which are based on known and experimentally observed neuroscience fail to explain certain critical aspects. These include a unitary sense of binding, non-computational aspects of conscious thinking, difference and transition between pre-conscious and conscious processing, (apparent) non-deterministic free will and the essential nature of our experience. We conclude that aspects of quantum theory (e.g. quantum coherence) and of a newly proposed physical phenomenon of wave function self-collapse (objective reduction, **OR**, [79]) offer possible solutions to each of these problematic features. We further conclude that cytoskeletal microtubules, which regulate intra-neuronal activities and have cylindrical paracrystalline structure, are the best candidates for sites of quantum action and **OR**, and of "orchestrated **OR**" (**Orch OR**). Accordingly, we present a model of consciousness based on the following assumptions:

- Coherent excitations (Fröhlich pumped phonons) among microtubule subunits (tubulins) support "cellular automaton" information processing in both classical (conformational) and quantum coherent superposition modes. Classical processing correlates with non-conscious, autonomic activity; quantum processing correlates with pre- and sub-conscious activity.
- The microtubule quantum coherent computing phase is able to be isolated from environmental interaction and maintain coherence for up to 500 msec (pre-conscious processing).
- A critical number of tubulins maintaining coherence within MTs for 500 msec collapses its own wave function (objective reduction: **OR**). This occurs because the mass–energy difference among the superpositioned states of coherent tubulins critically perturbs space–time geometry. To prevent multiple universes, the system must reduce to a single space–time by choosing eigenstates. The threshold for **OR** is related to quantum gravity; we calculate it in terms of the number of tubulins coherent for 500 msec to be very roughly 10^9 tubulins. Larger coherent sets will self-collapse faster, and smaller sets more slowly. Coherent sets which evolve over different time scales and in different brain distributions may be bound in an effectively simultaneous collapse which creates instantaneous "now". Cascades of these events constitute the familiar "stream of consciousness".
- Microtubule associated proteins (MAPs) and other tubulin modifications act as "nodes" to tune microtubule coherence and help to orchestrate collapse. We thus term the specific **OR** proposed to occur in microtubules and intrinsic to consciousness as "orchestrated objective reduction" (**Orch OR**).
- The **Orch OR** process, which introduces non-computability [78, 79], results in eigenstate patterns of tubulin conformational states which help direct neural function through the actions of microtubules.

In providing a connection among (1) pre-conscious to conscious transition, (2) fundamental space–time notions, (3) non-computability, and (4) binding of various (time scale and spatial) superpositions into instantaneous "now", we believe **Orch OR** in MTs is the most specific and plausible model for consciousness yet proposed.

Acknowledgements

SH is partially supported by NSF Grant No. DMS-9114503. We are grateful to Alwyn Scott and Richard Hofstad. Artwork by Dave Cantrell, Biomedical Communications, University of Arizona. Dedicated to Burnell R. Brown Jr. MD, Ph.D. FRCA. Reprinted in: *Toward a Science of Consciousness*: The First Tucson Discussions and Debates, S. Hameroff, A. Kaszniak, A. Scott, eds. (MIT Press, Cambridge, 1996).

References

[1] Y. Aharonov and L. Vaidman, Properties of a quantum system during the time interval between two measurements, Phys. Rev. A. 41 (1990) 11.
[2] L.A. Amos and A. Klug, Arrangement of subunits in flagellar microtubules, J. Cell Sci. 14 (1974) 523–550.
[3] J.M. Andreu, Hydrophobic interaction of tubulin, Ann. NY Acad. Sci. 466 (1986) 626–630.
[4] C. Aoki and P. Siekevitz, Ontogenic changes in the cyclic adenosine 3^1, 5^1 monophosphate-simulatable phosphorylation of cat visual cortex proteins, particularly of microtubule-associate protein 2 (MAP2): Effects of normal and dark rearing and of the exposure to light, J. Neurosci. 5 (1985) 2465–2483.
[5] J. Atema, Microtubule theory of sensory transduction, J. Theor. Biol. 38 (1973) 181–190.
[6] H. Athenstaedt, Pyroelectric and piezoelectric properties of vertebrates, Ann. NY Acad. Sci. 238 (1974) 68–93.
[7] B.J. Baars, A Cognitive Theory of Consciousness (Cambridge University Press, Cambridge, 1988).
[8] B.J. Baars, How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity?, in: Experimental and Theoretical Studies of Consciousness, Ciba Foundation Symposium, Vol. 174 (Wiley, Chichester, UK, 1993) 282–303.
[9] F. Beck and J.C. Eccles, Quantum aspects of brain activity and the role of consciousness, Proc. Nat. Acad. Sci. USA 89(23) (1992) 11357–11361.
[10] P. Benioff, Quantum mechanical Hamiltonian models of Turing machines, J. Stat. Phys. 29 (1982) 515–546.
[11] G. Bensimon and R. Chernat, Microtubule disruption and cognitive defects: Effect of colchicine on learning behavior in rats, Pharmacol. Biochem. Behavior 38 (1991) 141–145.
[12] D. Bigot and S.P. Hunt, Effect of excitatory amino acids on microtubule-associated proteins in cultured cortical and spinal neurons, Neurosci. Lett. 111 (1990) 275–280.
[13] R.B. Burns, Spatial organization of the microtubule associated proteins of reassembled brain microtubules, J. Ultrastruct. Res. 65 (1978) 73–82.
[14] H.F. Cantiello, J.L. Stow, A.G. Prat and D.A. Ausiello, Actin filaments regulate Na+ channel activity, Am. J. Physiol. 265(5 pt 1) (1991) C882–C888.
[15] K.-C. Chou, C.-T. Zhang and G.M. Maggiora, Solitary wave dynamics as a mechanism for explaining the internal motion during microtubule growth, Biopolymers 34 (1994) 143–153.
[16] C. Cianci, D. Graff, B. Gao and R.C. Weisenberg, ATP-dependent gelation-contraction of microtubules in vitro, Ann. NY Acad. Sci. 466 (1986) 656–659.
[17] J.S. Clegg, Intracellular water, metabolism and cell architecture, in: H. Fröhlich and F. Kremer, eds., Coherent Excitations in Biological Systems (Springer, Berlin, 1983).
[18] F.H.C. Crick, Function of the thalamic reticular complex: The searchlight hypothesis, Proc. Nat. Acad. Sci. USA 81 (1984) 4586–4593.
[19] F. Crick and C. Koch, Towards a neurobiological theory of consciousness, Seminars in the Neurosciences 2 (1990) 263–275.

[20] J. Cronley-Dillon, D. Carden and C. Birks, The possible involvement of brain microtubules in memory fixation, J. Exp. Biol. 61 (1974) 443–454.
[21] J.E. Dayhoff, S. Hameroff, R. Lahoz-Beltra and C.E. Swenberg, Cytoskeletal involvement in neuronal learning: A review, Eur. Biophys. J. 23 (1994) 79–93.
[22] M. De Brabander, A model for the microtubule organizing activity of the centrosomes and kinetochores in mammalian cells, Cell Biol. Intern. Rep. 6 (1982) 901–915.
[23] L. Deeke, B. Grotzinger and H.H. Kornhuber, Voluntary finger movement in man: Cerebral potentials and theory, Biol. Cybernetics 23 (1976) 99.
[24] E. Del Giudice, S. Doglia and M. Milani, Self-focusing and ponderomotive forces of coherent electric waves: A mechanisim for cytoskeleton formation and dynamics, in: H. Fröhlich and F. Kremer, eds., Coherent Excitations in Biological Systems (Springer, Berlin, 1983) 123–127.
[25] D. Deutsch, Quantum theory, the Church–Turing principle and the universal quantum computer, Proc. Roy. Soc. London Ser. A 400 (1985) 97–117.
[26] D. Deutsch and R. Josza, Rapid solution of problems by quantum computation, Proc. Roy. Soc. London Ser. A 439 (1992) 553–556.
[27] L. Diósi, Models for universal reduction of macroscopic quantum fluctuations, Phys. Rev. A. 40 (1989) 1165–1174.
[28] P. Dustin, Microtubules, 2nd Revised Edition (Springer, Berlin, 1984).
[29] J.C. Eccles, Evolution of consciousness, Proc. Nat. Acad. Sci. 89 (1992) 7320–7324.
[30] G. Edelman, The Remembered Present: A Biological Theory of Consciousness (Basic Books, New York, 1989).
[31] Y. Engelborghs, Dynamic aspects of the conformational states of tubulin and microtubules, Nanobiology 1 (1992) 97–105.
[32] R.P. Feynman, Quantum mechanical computers, Found. Phys. 16 (6) (1986) 507–531.
[33] W.J. Freeman, Mass Action in the Nervous System (Academic Press, New York, 1975).
[34] W.J. Freeman, Spatial properties of an EEG event in the olfactory bulb and cortex, Electroencephalogr. and Clin. Neurophysiol. 44 (1978) 586–605.
[35] P. Friedrich, Protein structure: The primary substrate for memory, Neurosci. 35 (1990) 1–7.
[36] H. Fröhlich, Long-range coherence and energy storage in biological systems. Int. J. Quantum Chem. 2 (1968) 641–649.
[37] H. Fröhlich, Long range coherence and the actions of enzymes, Nature 228 (1970) 1093.
[38] H. Fröhlich, The extraordinary dielectric properties of biological materials and the action of enzymes, Proc. Nat. Acad. Sci. 72 (1975) 4211–4215.
[39] L. Genberg, L. Richard, G. McLendon and R.J. Dwayne-Miller, Direct observation of global protein motion in hemoglobin and myoglobin on picosecond time scales, Science 251 (1991) 1051–1054.
[40] L. Genzel, F. Kremer, A. Poglitsch and G. Bechtold, Relaxation processes on a picosecond time scale in hemoglobin and poly observed by millimeter-wave spectroscopy, Biopolymers 22 (1983) 1715–1729.
[41] R. Geroch and J.B. Hartle, Computability and physical theories, Found. Phys. 16 (1986) 533.
[42] G.C. Ghirardi, R. Grassi and A. Rimini, Continuous-spontaneous reduction model involving gravity, Phys. Rev. A. 42 (1990) 1057–1064.
[43] G.C. Ghirardi, A. Rimini and T. Weber, Unified dynamics for microscopic and macroscopic systems, Phys. Rev. D. 34 (1986) 470.
[44] J.M. Gilbert and P. Strocchi, In vitro studies of the biosynthesis of brain tubulin, Ann. NY Acad. Sci. 466 (1986) 89–102.
[45] C.M. Gray and W. Singer, Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex, Proc. Nat. Acad. Sci. USA 86 (1989) 1698–1702.
[46] W. Grey-Walter, The Living Brain (Gerald Duckworth and Co., Ltd, 1953).
[47] W. Grundler and F. Keilmann, Sharp resonances in yeast growth prove nonthermal sensitivity to microwaves, Phys. Rev. Lett. 51 (1983) 1214–1216.
[48] S. Halpain and P. Greengard, Activation of NMDA receptors induces rapid dephosphorylation of the cytoskeletal protein MAP2, Neuron 5, 237–246.
[49] S.R. Hameroff, Ultimate Computing: Biomolecular Consciousness and Nanotechnology (North-Holland, Amsterdam, 1987).

[50] S.R. Hameroff, J.E. Dayhoff, R. Lahoz-Beltra, A. Samsonovich and S. Rasmussen, Conformational automata in the cytoskeleton: Models for molecular computation, IEEE Comput. (1992) 30–39.

[51] S.R. Hameroff, S.A. Smith and R.C. Watt, Nonlinear electrodynamics in cytoskeletal protein lattices, in: W.R. Adey and A.F. Lawrence, eds., Nonlinear Electrodynamics in Biological Systems (Plenum, New York, 1984).

[52] S.R. Hameroff and R.C. Watt, Information processing in microtubules, J. Theor. Biol. 98 (1982) 549–561.

[53] D.O. Hebb, The Organization of Behavior (Wiley, New York, 1949).

[54] D.O. Hebb, Essay on Mind (Lawrence Erlbaum Associates, Hillsdale, NJ, 1980).

[55] N. Hirokawa, Molecular architecture and dynamics of the neuronal cytoskeleton, in: R.D. Burgoyne, ed., The Neuronal Cytoskeleton (Wiley-Liss, New York, 1991) 5–74.

[56] M. Jibu, S. Hagan, S.R. Hameroff, K.H. Pribram and K. Yasue, Quantum optical coherence in cytoskeletal microtubules: Implications for brain function, BioSystems 32 (1994) 195–209.

[57] A. Kaivarainen, Personal communication, 1994.

[58] F. Károlházy, A. Frenkel and B. Lukacs, On the possible role of gravity on the reduction of the wave function, in : R. Penrose and C.J. Isham, eds., Quantum Concepts in Space and Time (Oxford University Press, Oxford, 1986).

[59] M. Karplus and J.A. McCammon, Protein ion channels, gates, receptors, in: J. King, ed., Dynamics of Proteins: Elements and Function, Ann. Rev. Biochem. (Benjamin/Cummings, Menlo Park, 1983) 263–300.

[60] H. Kim, C.G. Jensen and L.I. Rebhund, The binding of MAP2 and tau on brain microtubules in vitro, in: D. Soifer, ed., Dynamic Aspects of Microtubule Biology (1986) 218–239; Ann. NY Acad. Sci. 466.

[61] M. Kirschner and T. Mitchison, Beyond self assembly: From microtubules to morphogenesis, Cell 45 (1986) 329–342.

[62] T. Kudo, K. Tada, M. Takeda and T. Nishimura, Learning impairment and microtubule-associated protein 2 (MAP2) decrease in gerbils under chronic cerebral hypoperfusion, Stroke 21 (1990) 1205–1209.

[63] J.C. Lee, D.J. Field, H.J. George and J. Head, Biochemical and chemical properties of tubulin subspecies, Ann. NY Acad. Sci. 466 (1986) 111–128.

[64] B. Libet, Cerebral processes that distinguish conscious experience from unconscious mental functions, in: J.C. Eccles and O.D. Creutzfeld, eds., The Principles of Design and Operation of the Brain (Springer, Berlin, 1990) 185–205; Experimental Brain Research Series 21.

[65] B. Libet, E.W. Wright, Jr., B. Feinstein and D.K. Pearl, Subjective referral of the timing for a conscious sensory experience, Brain 102 (1979) 193–224.

[66] E. Mandelkow and E.-M. Mandelkow, Microtubule structure, Curr. Opinions Structural Biology 4 (1994) 171–179.

[67] E.-M. Mandelkow and E. Mandelkow, Tau as a marker for Alzheimer's disease, Trends Biol. Sci. 18 (1993) 480–483.

[68] I.N. Marshall., Consciousness and Bose–Einstein condensates, New ideas in Psychology 7 (1989) 73–83.

[69] S. Mascarenhas, The electret effect in bone and biopolymers and the bound water problem, Ann. NY Acad. Sci. 238 (1974) 36–52.

[70] S.S. Matsuyama and L.F. Jarvik, Hypothesis: Microtubules, a key to Alzheimer's disease, Proc. Nat. Acad. Sci. 86 (1989) 8152–8156.

[71] R. Melki, M.F. Carlier, D. Pantaloni and S.N. Timasheff, Cold depolymerization of microtubules to double rings: Geometric stabilization of assemblies, Biochemistry 28 (1989) 9143–9152.

[72] R. Mileusnic, S.P. Rose and P. Tillson, Passive avoidance learning results in region specific changes in concentration of, and incorporation into, colchicine binding proteins in the chick forebrain, Neur. Chem. 34 (1980) 1007–1015.

[73] R.J. Montoro, J. Diaz-Nido, J. Avila and J. Lopez-Barneo, N-methyl-d-aspartate stimulates the dephosphorylation of the microtubule-associated protein 2 and potentiates excitatory synaptic pathways in the rat hippocampus, Neuroscience 54 (4) (1993) 859–871.

[74] C. Neubauer, A.M. Phelan, H. Keus and D.G. Lange, Microwave irradiation of rats at 2.45 GHz activates pinocytotic-like uptake of tracer by capillary endothelial cells of cerebral cortex, Bioelectromagnetics 11 (1990) 261–268.

[75] P. Pearle, Combining stochastic dynamical state vector reduction with spontaneous localization, Phys. Rev. D. 13 (1989) 857–868.

[76] P. Pearle and E. Squires, Bound-state excitation, nucleon decay experiments and models of wave-function collapse, Phys. Rev. Lett. 73 (1) (1994) 1–5.

[77] R. Penrose, Newton, quantum theory and reality, in: S.W. Hawking and W. Israel, eds., 300 Years of Gravity (Cambridge University Press, Cambridge, 1987).

[78] R. Penrose, The Emperor's New Mind (Oxford University Press, Oxford, 1989).
[79] R. Penrose, Shadows of the Mind (Oxford University Press, London, 1994).
[80] R. Penrose and L. Onsager, Bose–Einstein condensation and liquid helium, Phys. Rev. 104 (1956) 576–584.
[81] T.T. Puck and A. Krystosek, Role of the cytoskeleton in genome regulation and cancer, Int. Rev. Cytology 132 (1992) 75–108.
[82] S. Rasmussen, H. Karampurwala, R. Vaidyanath, K.S. Jensen and S. Hameroff, Computational connectionism within neurons: A model of cytoskeletal automata subserving neural networks, Phys. D 42 (1990) 428–449.
[83] L.E. Roth and D.J. Pihlaja, Gradionation: Hypothesis for positioning and patterning, J. Protozoology 24 (1) (1977) 2–9.
[84] M.V. Sataric, R.B. Zakula and J.A. Tuszynski, A model of the energy transfer mechanisms in microtubules involving a single soliton, Nanobiology (1992) 445–456.
[85] P. Satir, Cytoplasmic matrix: Old and new questions, J. Cell Biol. 99 (1) (1984) 235–238.
[86] A.C. Scott, Stairway to the Mind (Springer, Berlin, 1995).
[87] W. Singer, Synchronization of cortical activity and its putative role in information processing and learning, Ann. Rev. Physiol. 55 (1993) 349–374.
[88] H.P. Stapp, Mind, Matter and Quantum Mechanics (Springer, Berlin, 1993).
[89] W.E. Theurkauf and R.B. Vallee, Extensive cAMP-dependent and cAMP-independent phosphorylation of microtubule associated protein 2, J. Biol. Chem. 258 (1983) 7883–7886.
[90] J. Tuszynski, J., S. Hameroff, M.V. Sataric, B. Trpisova and M.L.A. Nip, Ferroelectric behavior in microtubule dipole lattices; Implications for information processing, signaling and assembly/disassembly, J. Theor. Biol. (1995) in press.
[91] P. Vassilev, M. Kanazirska and H.T. Tien, Intermembrane linkage mediated by tubulin, Biochem. Biophys. Res. Comm. 126 (1985) 559–565.
[92] C. Von der Marlsburg and W. Schneider, A neural cocktail party processor, Biol. Cybern. 54 (1986) 29–40.
[93] J. Von Neumann, Theory of self-reproducing automata, in: A.W. Burks, ed. (University of Illinois Press, Urbana, 1966).
[94] N. Wang and D.E. Ingber, Control of cytoskeletal mechanics by extracellular matrix, cell shape and mechanical tension, Biophysical J. 66 (6) (1994) 2181–2189.
[95] N.J. Woolf, S.L. Young, G.V.W. Johnson and M.S. Fanselow, Pavlovian conditioning alters cortical microtubule-associated protein-2, NeuroReport 5 (1994) 1045–1048.
[96] W. Yu and P.W. Baas, Changes in microtubule number and length during axon differentiation, J. Neuroscience 14 (5) (1994) 2818–2829.

248

SPACE-TIME AND COSMOLOGY

The Tanner lectures on human values, Vol. 17. Ed. G. B. Peterson, 347–91. University of Utah Press, 1996

712 Space-time and cosmology. *The Tanner lectures on human values*, Vol. 17. Ed. G. B. Peterson, 347–91. 1996

Space-time and Cosmology

ROGER PENROSE

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

Cambridge University
February 13–15, 1995

SIR ROGER PENROSE is Rouse Ball Professor of Mathematics at Oxford University, as well as Francis and Helen Pentz Distinguished Professor of Physics and Mathematics at Pennsylvania State University. He was educated at University College, London, and received his Ph.D. from Cambridge University. He is a fellow of the Royal Society and recipient of numerous prizes and awards, including the 1988 Wolf Prize that he shared with Steven Hawking for their understanding of the universe. His numerous publications include *Shadows of the Mind* (1994), *The Emperor's New Mind* (1994), *Spinors and Space Time* (1984 and 1986, coauthor), and *Techniques of Differential Topology* (1972). He was knighted in 1994 for services to science.

The title of these lectures is "The Large, the Small, and the Human Mind" and the subject of this first lecture is the Large.[1] The first and second lectures are concerned with our physical Universe, which I represent very schematically as the "sphere" in my first picture.

However, these will not be "botanical" lectures, telling you in detail what is here and what is there in our Universe, but rather I want to concentrate upon understanding of the actual laws that govern the way the world behaves. One of the reasons that I have chosen to divide my descriptions of the physical laws between two lectures, namely, the Large and the Small, is that the laws that govern the large-scale behaviour of the world and those that govern its small-scale behaviour seem to be very different. The fact that they seem to be so different, and what we might have to do about his seeming discrepancy, is central to the subject of the third lecture — which is where the human mind comes in.

Since I shall be talking about the physical world in terms of the physical theories that underlie its behaviour, I shall also have to say something about another world, the Platonic world of absolutes, in its particular role as the world of mathematical truth.

Physical
World

FIGURE 1.

---

[1] This is the first of three lectures with the overall title "The Large, the Small, and the Human Mind." All three lectures will be published by Cambridge University Press.

One can well take the view that the "Platonic world" contains other absolutes, such as the Good and the Beautiful, but I shall be concerned here only with the Platonic concepts of mathematics. Some people find it hard to conceive of this world as existing on its own. They may prefer to think of mathematical concepts merely as idealisations of our physical world — and, on this view, the mathematical world would be thought of as emerging from the world of physical objects (figure 2).

Now, this is not how I think of mathematics, nor, I believe, is it how most mathematicians or mathematical physicists think about the world. They think about it in a rather different way, as a structure precisely governed according to timeless mathematical laws. Thus, they prefer to think of the physical world, more appropriately, as emerging out of the ("timeless") world of mathematics, as illustrated in figure 3. This picture will have importance for what I shall say in the third lecture, and it also underlies most of what I shall say in the first two.

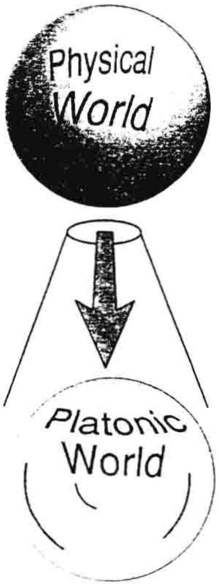


FIGURE 2.

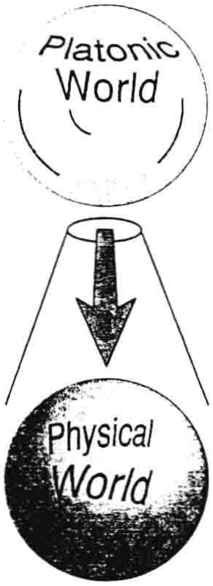


FIGURE 3.

One of the remarkable things about the behaviour of the world is how it seems to be grounded in mathematics to a quite extraordinary degree of accuracy. The more we understand about the physical world, and the deeper we probe into the laws of nature, the more it seems as though the physical world almost evaporates and we are left only with mathematics. The more deeply we understand the laws of physics, the more we are driven into this world of mathematics and of mathematical concepts.

Let us look at the scales we have to deal with in the Universe and also the role of our place in the Universe. I can summarise all these scales in a single diagram (figure 4). On the left-hand side

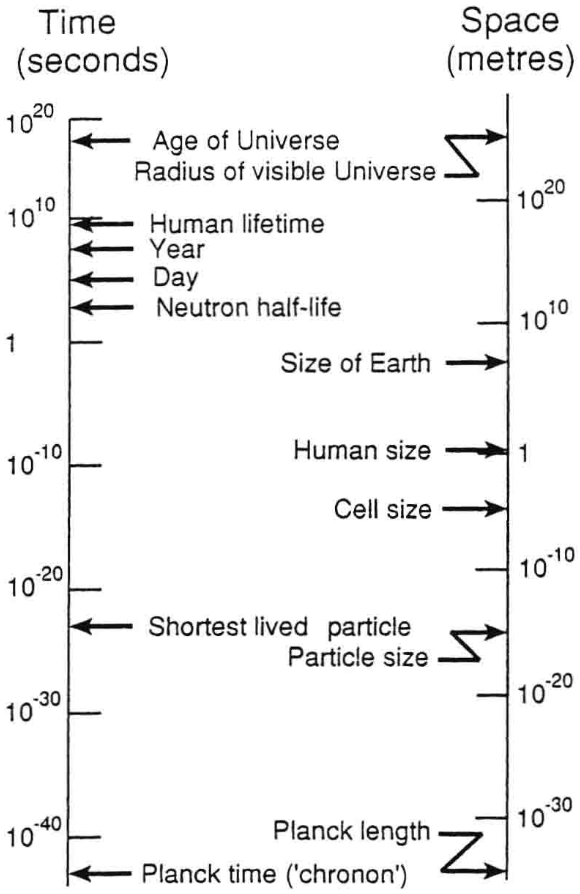

Time
(seconds)
Space
(metres)
10^{20}
Age of Universe
Radius of visible Universe
10^{20}
10^{10}
Human lifetime
Year
Day
Neutron half-life
10^{10}
1
Size of Earth
10^{-10}
Human size
1
Cell size
10^{-10}
10^{-20}
Shortest lived particle
Particle size
10^{-20}
10^{-30}
10^{-30}
10^{-40}
Planck length
Planck time ('chronon')

FIGURE 4. Sizes and time-scales in the Universe.

of the diagram, time-scales are shown. At the bottom of the diagram, on the left-hand side, is the very shortest time scale that is physically meaningful. This time-scale is about 10^{-43} of a second and is often referred to as the *Planck time-scale* or a "chronon." This time-scale is much shorter than anything experienced in particle physics. For example, the shortest-lived particles, called resonances, last for about 10^{-23} of a second. Further up the diagram, on the left, the day and the year are shown, and, at the top of the diagram, the present age of the Universe is shown.

On the right-hand side of the diagram, distances corresponding to these time-scales are depicted. The length corresponding to the Planck time (or chronon) is the fundamental unit of length, called the *Planck length*. These concepts of the Planck time and the Planck length fall out naturally when one tries to combine the physical theories that describe the large and the small, that is, combining Einstein's General Relativity, which describes the physics of the very large, with quantum mechanics, which describes the physics of the very small. When these theories are brought together, these Planck lengths and times turn out to be fundamental. The translation from the left-hand to the right-hand axis of the diagram is via the speed of light so that times can be translated into distances by asking how far a light signal could travel in that time.

The sizes of the physical objects represented on the diagram range from about 10^{-15} of a metre for the characteristic sizes of particles to about 10^{27} meters for the radius of the observable Universe at the present time, which is roughly the age of the Universe multiplied by the speed of light. It is intriguing to note where *we* are in the diagram, namely, the human scale. With regard to spatial dimensions, it can be seen that we are more or less in the middle of the diagram. We are enormous compared with the Planck length; even compared with the size of particles, we are very large. Yet, compared with the distance scale of the observable Universe, we are very tiny. Indeed, we are as small compared with it as we are large compared with particles. In contrast, with regard to temporal dimensions, the human lifetime is almost as long as the Universe! People talk about the ephemeral nature of existence but, when you look at the human lifetime as shown in the diagram, it can be seen that we are not ephemeral at all — we live more or less as long as the Universe itself! Of course, this is looking on a "logarithmic scale," but this is the natural thing to do when we are concerned with such enormous ranges. To put it another way, the number of human lifetimes that make up the age of the Universe is very, very much less than the number of Planck times, or even lifetimes of the shortest-lived particles, that make up a human lifetime. Thus, we are really very stable structures in the Universe. As far as spatial sizes are concerned, we are very much in the middle — we directly experience neither the physics of the very large nor the physics of the very small. We are very much in-between. In fact, looked at logarithmically, all living objects from single cells to human beings are roughly the same in-between size.

What kinds of physics apply on these different scales? Let me introduce the diagram that summarizes the whole of physics. I have had to leave out a few details, of course, such as all the equations! But the essential basic theories that physicists use are indicated.

The key point is that, in physics, we use two very different types of procedure. To describe the small-scale behaviour, we use quantum mechanics — what I have described as the quantum level in figure 5. I shall say much more about this in the second lecture. One of the things that people say about quantum mechanics is that it is fuzzy and indeterministic, but this is not true. So long as you remain at this level, quantum theory is deterministic and precise. In its most familiar form, quantum mechanics involves use of the equation known as Schrödinger's equation, which governs the behaviour of the physical state of a quantum system — called its *quantum state* — and this is a deterministic equation. I have used the letter *U* to describe this quantum level activity. Indeterminacy

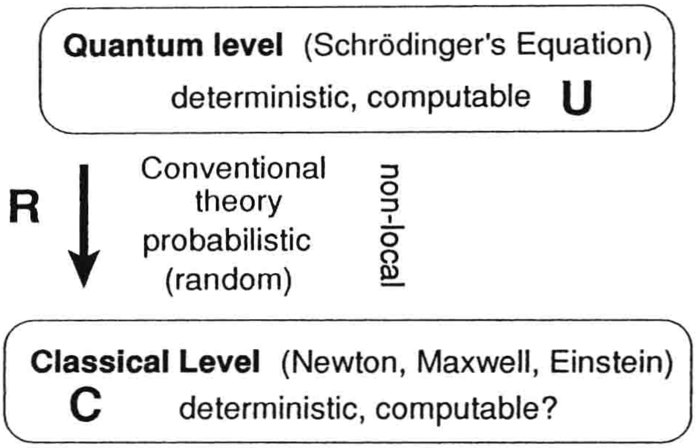

Quantum level (Schrödinger's Equation)
deterministic, computable **U**
R
Conventional theory probabilistic (random)
non-local
Classical Level (Newton, Maxwell, Einstein)
C deterministic, computable?

FIGURE 5.

in quantum mechanics arises only when you perform what is called "making a measurement" and that involves magnifying an event from the quantum level to the classical level. I shall say quite a lot about this in the second lecture.

On the large scale, we use classical physics, which is entirely deterministic — these classical laws include Newton's laws of motion, Maxwell's laws for the electromagnetic field, which incorporate electricity, magnetism, and light, and Einstein's theories of relativity, the Special Theory, which deals with large velocities, and the General Theory, which deals with large gravitational fields. These laws apply very, very accurately on the large scale.

Just as a footnote to figure 5, it can be seen that I have included a remark about "computability" in quantum and classical physics. This has no relevance to the present lecture or the next, but it will have importance for the third, and I shall return to the issue of computability in that lecture.

For the rest of the present lecture, I shall be primarily concerned with Einstein's theory of relativity—specifically, how theory works, its extraordinary accuracy, and something about its elegance as a physical theory. But let us first consider Newtonian theory. Newtonian physics, just as in the case of relativity, allows a space-time description to be used. This was first precisely formulated by E.-J. Cartan for Newtonian gravity, some time after Einstein had presented his General Theory of relativity. The physics of Galileo and Newton is represented in space-time for which there is a global time coordinate, here depicted as running up the diagram (figure 6); and for each constant value of the time, there is a space section which is a Euclidean 3-space, here depicted as horizontal planes. An essential feature of the Newtonian space-time picture is that these space-slices, across the diagram, represent moments of simultaneity.

Thus, everything that occurs on Monday at noon lies on one horizontal slice through the space-time diagram; everything that happens on Tuesday at noon lies on the next slice shown in the

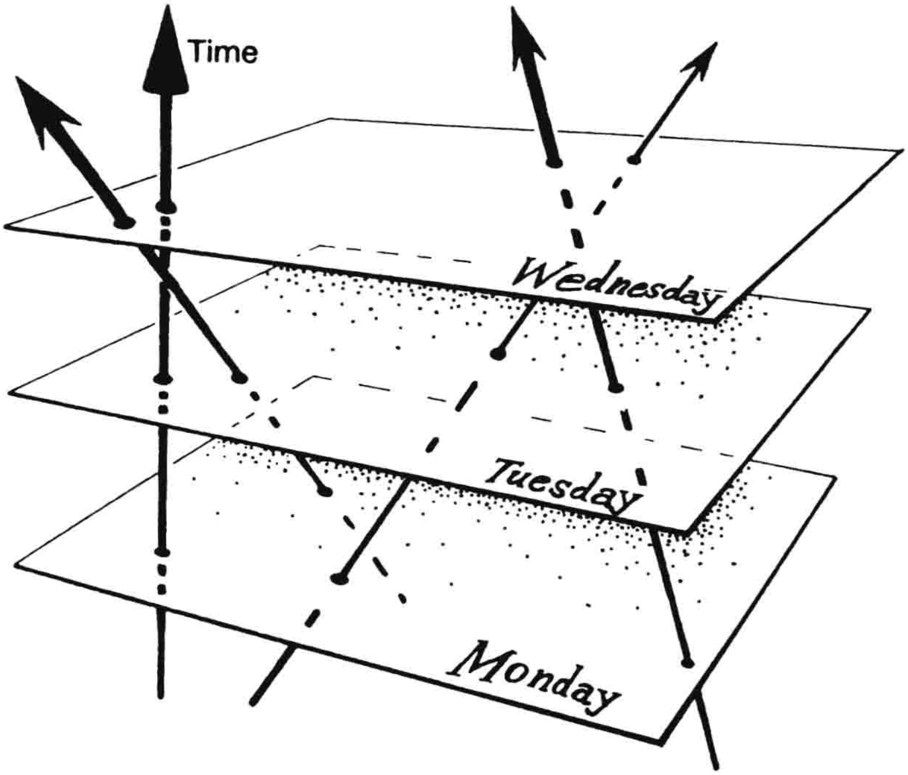


FIGURE 6. Galilean space-time: particles in uniform motion are depicted as straight lines.

diagram; and so on. Time cuts across the space-time diagram and the Euclidean sections follow one after the other as time progresses. All the observers can agree about the time when events take place, no matter how they move through the space-time, because everyone uses the same time slices to measure how time passes.

In Einstein's Special Theory of relativity, one has to adopt a different picture. In it, the space-time picture is absolutely essential — the key difference is that time is not the universal thing it is in Newtonian theory. To appreciate how the theories differ, it is necessary to understand an essential part of relativity theory, namely, those structures known as *light-cones*.

What is a light-cone? A light-cone is drawn in figure 7. We imagine a flash of light taking place at some point at some instant — that is, at an *event* in space-time — and the light waves travel outwards from this event, the source of the flash, at the speed of light. In a purely spatial picture (right-hand picture of figure 7), we can represent the paths of the light waves through space as a sphere expanding at the speed of light. We can now translate this motion of the light waves into a space-time diagram (left-hand picture of figure 7) in which time runs up the diagram and the space coordinates refer to horizontal displacements, just as

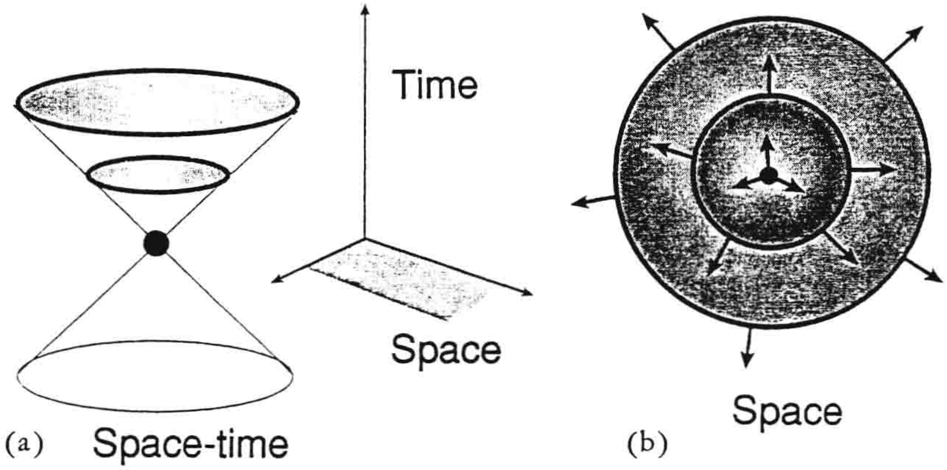


FIGURE 7. The representation of the history of a light flash in terms of its propagation in (a) space-time and (b) space.

in the Newtonian situation of figure 6. Unfortunately, in the full space-time picture, on the left of figure 7, we can represent only two spatial dimensions horizontally on the diagram, because the space of our picture is only three-dimensional. Now, we see that the flash is represented by a point (event) at the origin and that the subsequent paths of the light rays (waves) cut the horizontal "space" planes in circles, the radii of which increase at the speed of light up the diagram. It can be seen that the paths of the light rays form cones in the space-time diagram. The light-cone thus represents the history of this flash of light — light propagates away from the origin along the light-cone, which means at the speed of light, into the future. Light rays can also arrive at the origin along the light-cone from the past — that part of the light-cone is known as the past light-cone and all information carried to the observer by light waves arrives at the origin along this cone.

Light-cones represent the most important structures in space-time. In particular, they represent the limits of causal influence. The history of a particle in space-time is represented by a line travelling up the space-time diagram, and this line has to lie within the light-cone (figure 8). This is just another way of saying that a

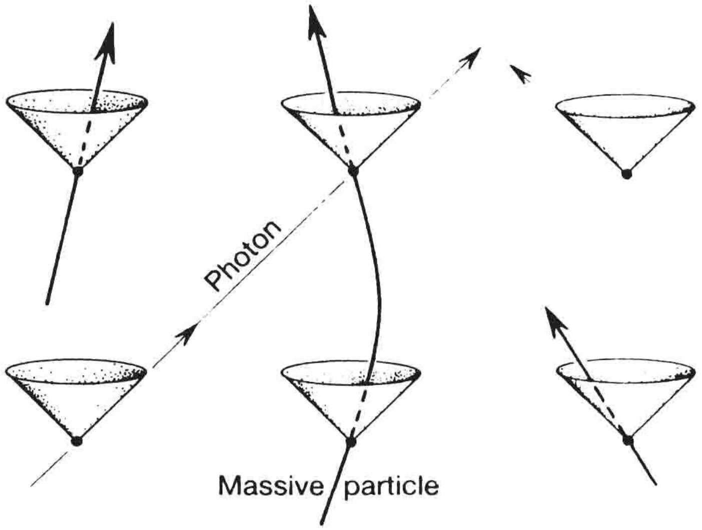


FIGURE 8. A picture of Minkowski geometry.

material particle cannot travel faster than the speed of light. No signal can travel from inside to outside the future light-cone, so that the light-cone does indeed represent the limits of causality.

There are some remarkable geometrical properties that relate to the light-cones. Let us consider two observers moving at different speeds through space-time. Unlike the case of Newtonian theory, in which the planes of simultaneity are the same for all observers, there is no absolute simultaneity in relativity. Observers moving at different speeds draw their own planes of simultaneity as different sections through space-time, as illustrated in figure 9. There is a very well-defined way of transforming from one plane to another through what is known as a *Lorentz transformation*,

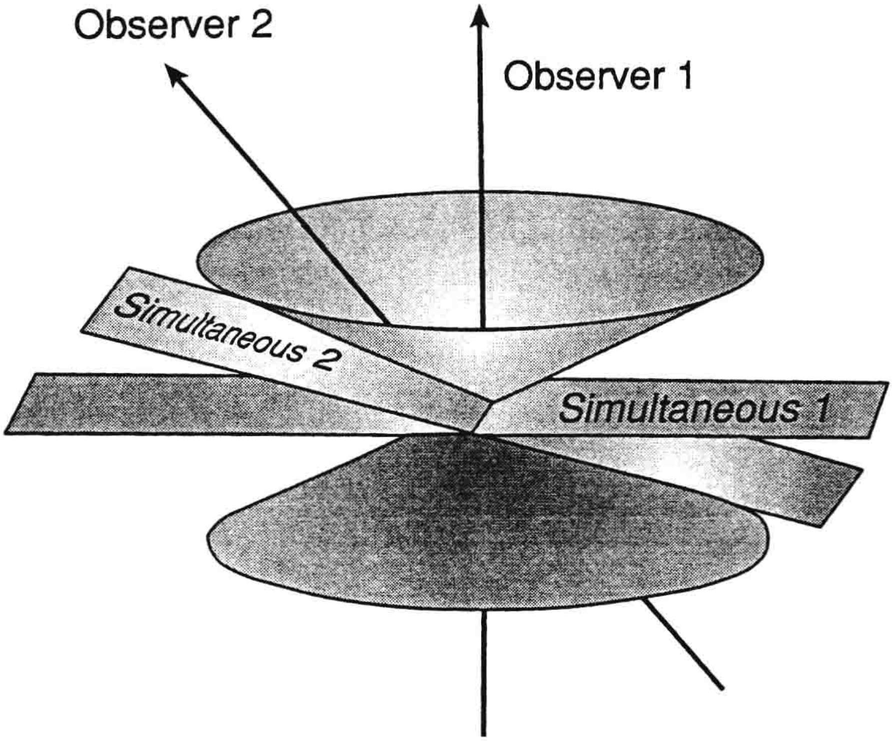


FIGURE 9. Illustrating the relativity of simultaneity according to Einstein's Special Theory of relativity. Observers 1 and 2 are moving relative to one another through space-time. Events that are simultaneous for Observer 1 are not simultaneous for Observer 2 and vice versa.

these transformations constituting what is called the *Lorentz group*. The discovery of this group was an essential feature in the discovery of Einstein's Special Theory of relativity. The Lorentz group can be understood as a group of (linear) space-time transformations, leaving a light-cone invariant.

We can also appreciate the Lorentz group from a slightly different viewpoint. As I have emphasised, the light-cones are the fundamental structures of space-time. Imagine that you are an observer located somewhere in space, looking out at the Universe. What you see are the light rays coming from the stars to your eyes. According to the space-time viewpoint, the events you observe are the intersections of the world-lines of the stars with your past light-cone, as illustrated in figure 10(a). You observe along your past light-cone the positions of the stars at particular points. These points seem to be situated on the celestial sphere that appears to surround you. Now, imagine another observer, moving at some great speed relative to you, who passes closely by you at the moment you both look out at the sky. This second observer perceives the same stars as you do, but finds them to be located in different positions on the celestial sphere — this is the effect known as *aberration*. There is a set of transformations that enables us to work out the relationship between what each of these observers sees on his or her celestial sphere. Each of these transformations is one that takes a sphere to a sphere. But it is one of a very special kind. It takes exact circles to exact circles and it preserves angles. Thus, if a pattern in the sky appears to be circular to you, then it must also appear circular to the other observer.

There is a very beautiful way of describing how this works and I illustrate it to show that there is a particular elegance in the mathematics that often underlies physics at its most fundamental level. Figure 10(c) shows a sphere with a plane drawn through its equator. We can draw figures on the surface of the sphere and then examine how they are projected to the equatorial plane from the south pole, as illustrated. This type of projection is known as

Space-time and cosmology. *The Tanner lectures on human values*, Vol. 17. Ed. G. B. Peterson, 347–91. 1996

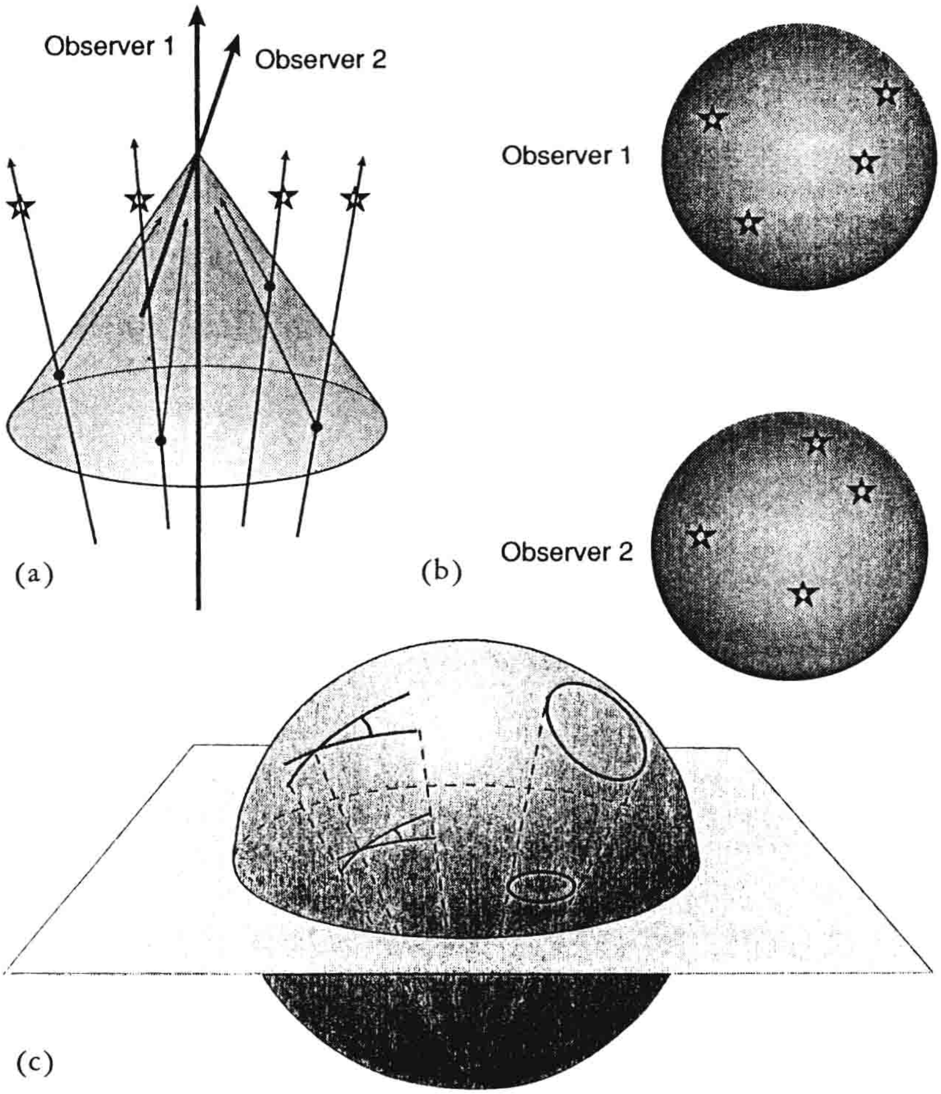


FIGURE 10. Illustrating how observations are made of the sky by Observers 1 and 2. (a) Observer 1 observes stars along the past light-cone. The points at which stars cross the light-cone are indicated by black dots. Light signals propagate from the stars to the observers along the light-cone as illustrated. Observer 2 is moving through space-time at a certain speed relative to Observer 1. (b) Illustrating the location of stars on the sky as observed by Observer 1 and Observer 2, when they are coincident at some point in space-time. (c) Illustrating the use of the Riemann sphere to locate the positions of stars on the celestial sphere for observers moving at a constant speed relative to each other.

a stereographic projection and it has some rather extraordinary properties. Circles on the sphere are projected into exact circles on the plane, and the angles between curves on the sphere are projected into exactly the same angles on the plane. As I shall discuss more fully in the second lecture, this projection allows us to label the points of the sphere by complex numbers (numbers involving the square root of -1) that can be used to label the points of the equatorial plane, together with "infinity," to give it the structure known as the "Riemann sphere."

For those who are interested, the transformation is

$$\mu \rightarrow \mu' = \frac{\alpha\mu + \beta}{\lambda\mu + \delta}$$

It is a fact well known to mathematicians that this transformation sends circles into circles and preserves angles. Transformations of this kind are known as Möbius transformations. For our present purposes, we need merely note the simple elegance of the form of the Lorentz (aberration) formula when written in terms of such a complex parameter μ.

A striking point about this way of looking at these transformations is that, according to Special Relativity, the formula is very simple, whereas, in expressing the corresponding aberration transformation according to Newtonian Mechanics, the formula would be much more complicated. It often turns out that, when you get down to the fundamentals and develop a more exact theory, the mathematics turns out to be simpler, even if the formalism appears to be more complicated in the first instance. This important point is exemplified by the contrast between Galilean and Einstein's relativity.

Thus, in the Special Theory of relativity, we have a theory that is, in many ways, simpler than Newtonian mechanics. From the point of view of mathematics, and particularly from the point of view of group theory, it is a much nicer structure. In Special Relativity, the space-time is flat and all the light cones are lined up regularly, as illustrated in figure 8. If we now go one step further to Einstein's General Relativity, that is, the theory of space-time in the presence of gravity, the picture seems at first sight rather muddied up — the light-cones are all over the place (figure 11). Now, I have been saying that, as we develop deeper and deeper theories, the mathematics becomes simpler, but look what has happened here — I had a nice elegant piece of mathematics that has become horribly complicated. Well, that sort of thing happens — you will have to bear with me for a little while until the simplicity reappears.

Let me remind you of the fundamental ingredients of Einstein's theory of gravity. One basic ingredient is called Galileo's Principle of Equivalence. In figure 12, I show Galileo leaning over from the top of the Tower of Pisa dropping large and small rocks. Whether or not he actually performed this experiment, he certainly well understood that, if the effects of air resistance are ignored, the two rocks would fall to the ground in the same time. If you happened to be sitting on one of these rocks looking at the other one as they fall together, you would observe the other rock hovering in front of you (I have shown a camcorder attached to one of the rocks to make the observation). Nowadays, with space travel, this is a very familiar phenomenon — just recently, we have seen a British-born astronaut walking in space, and, just like the big rock and the little rock, the spaceship hovers in front of the astronaut — this is exactly the same phenomenon as Galileo's Principle of Equivalence.

Thus, if you look at gravity in the right way, that is, in a falling frame of reference, it seems to disappear right in front of your eyes. This is indeed correct. But Einstein's theory does *not* tell you that gravity disappears — it only tells you that the *force* of gravity disappears. There is something left and that is the tidal effect of gravity.

Let me introduce a little bit more mathematics, but not much. We need to describe the curvature of space-time and this is de-

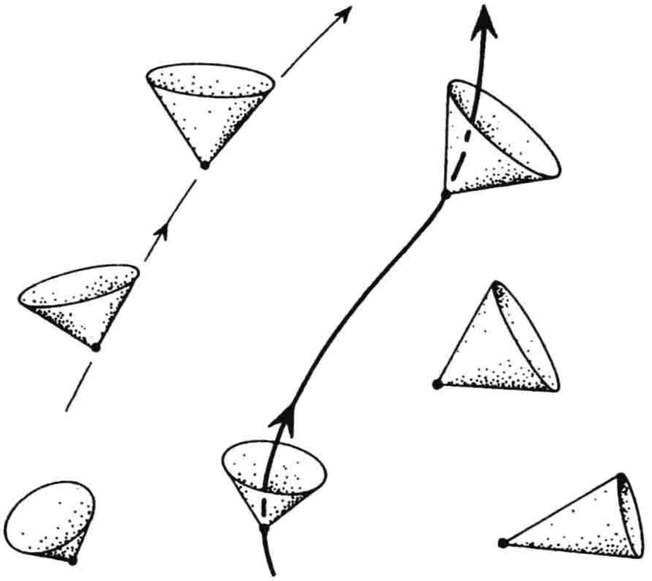


FIGURE 11. A picture of curved space-time.

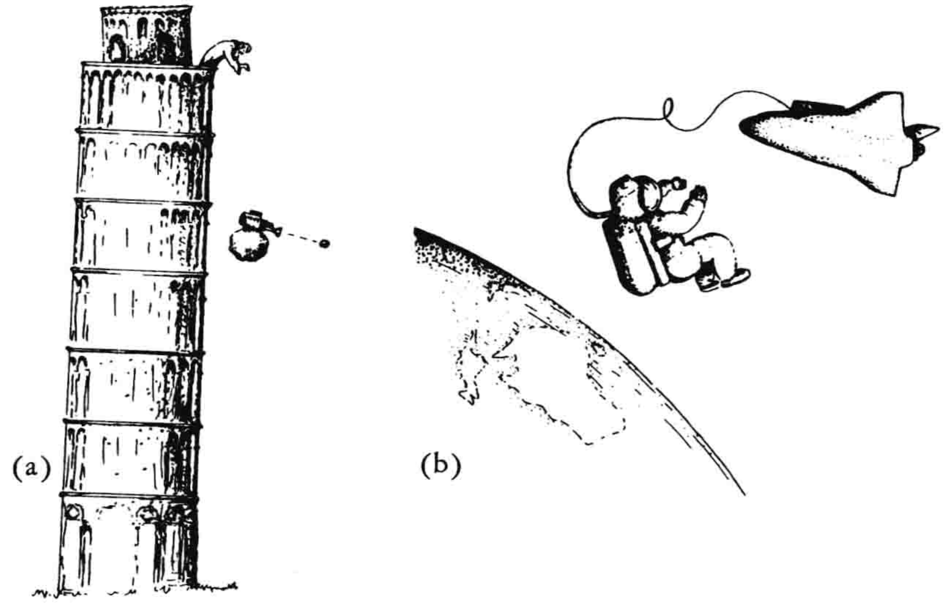


FIGURE 12. (a) Galileo dropping two rocks (and a camcorder) from the Leaning Tower of Pisa. (b) The astronaut sees the space-vehicle hover before him or her, seemingly unaffected by gravity.

scribed by an object known as a tensor which I have called *Riemann* in the following equation. It is actually called the Riemann curvature tensor but I will not tell you what it is except that it is represented by a capital *R* with a number of indices stuck on the bottom, which are indicated by the dots. The Riemann curvature tensor is made up of two pieces. One of the pieces is called the *Weyl* curvature and the other piece is called the *Ricci* curvature, and we have the (schematic) equation

$$\textit{Riemann} = \textit{Weyl} + \textit{Ricci}$$
$$R\,.\,.\,.\,. = C\,.\,.\,.\,. + R\,.\,.\,g\,.\,.$$

Formally, *C*. . . . and *R*. . are the Weyl and Ricci curvature tensors, respectively, and *g*. . is the metric tensor.

The Weyl curvature measures the tidal effect. What is the "tidal" effect? Recall that, from the astronaut's point of view, it seems that gravity has been abolished, but that is not quite true. Imagine that the astronaut is surrounded by a sphere of particles, which are initially at rest with respect to the astronaut. Now, at first they will just hover there but soon they will start to accelerate because of the slight differences in the gravitational attraction of the Earth at different points in the sphere. (Notice that I am describing the effect in Newtonian language, but that is quite adequate.) These slight differences cause the original sphere of particles to become distorted into an elliptical arrangement, as illustrated in figure 13(a).

This distortion occurs partly because of the slightly greater attraction of the Earth for those particles closer to the Earth and lesser attraction for those further away, and partly because, at the sides of the sphere, the Earth's attraction acts slightly inwards. This causes the sphere to be distorted into an ellipsoid. It is called the tidal effect for the very good reason that, if you replace the Earth by the Moon and the sphere of particles by the Earth with its oceans, then the effect of the Moon on the surface of the oceans is

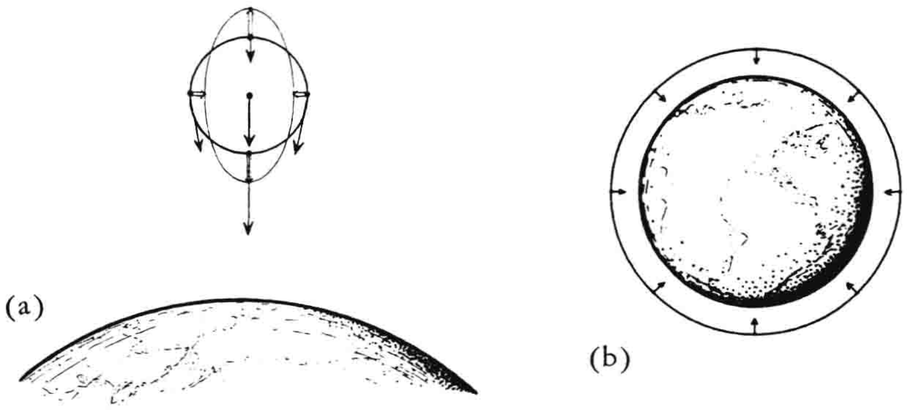


FIGURE 13. (a) The tidal effect. Double arrows show relative acceleration. (b) When the sphere surrounds matter (here on Earth), there is a net inward acceleration.

exactly the gravitational effect upon the sphere of particles — the sea surface closest to the Moon is pulled towards it, whereas the ocean surface on the other side of the Earth is, in effect, pushed away from it. The effect causes the sea surface to bulge out on either side of the Earth and is the cause of the two high tides that occur each day.

The effects of gravity, from Einstein's point of view, are simply this tidal effect. It is defined precisely by the Weyl curvature, that is, the part *C*. . . . of the Riemann curvature. This part of the curvature tensor is volume-preserving — that is, if you work out the initial accelerations of the particles of the sphere, the volume of the sphere and the volume of the ellipsoid into which it is distorted are initially the same.

The remaining part of the curvature is known as the *Ricci* curvature and it has a volume-reducing effect. From figure 13(b), it can be seen that if, instead of being at the bottom of the diagram, the Earth were inside the sphere of particles, the volume of the sphere of particles would be reduced as the particles accelerate inwards. The amount of this reduction in volume is a measure of the Ricci curvature. Einstein's theory tells us that the Ricci curvature is determined by the amount of matter present within a small sphere about that point in space. In other words, the density of matter, appropriately defined, tells us how the particles are accelerated inwards at that point in space. Einstein's theory is almost the same as Newton's when expressed in this way.

This is how Einstein formulates his theory of gravity — it is expressed in terms of the tidal effects, which are measurements of the local space-time curvature. It is crucial that we have to think in terms of the curvature of four-dimensional space-time. This was shown schematically in figure 11 — we think of the lines that represent the world lines of particles and the ways in which these paths are distorted as a measurement of the curvature of space-time. Thus, Einstein's theory is essentially a geometric theory of four-dimensional space-time — it is an extraordinarily beautiful theory mathematically.

The history of Einstein's discovery of the theory of General Relativity contains an important moral. It was first fully formulated in 1915. It was not motivated by any observational need but by various aesthetic, geometric, and physical desiderata. The key ingredients were Galileo's Principle of Equivalence, exemplified by his dropping rocks of different masses (figure 12), and the ideas of non-Euclidean geometry, which is the natural language for describing the curvature of space-time. There was not a great deal on the observational side in 1915. Once General Relativity was formulated in its final form, it was realised that there were the three key observational tests of the theory. The perihelion of the orbit of Mercury is swung around, or precesses, in a way that could not be explained by the Newtonian gravitational influence of the other planets — General Relativity predicts exactly the observed precession. The paths of light rays are bent by the Sun and this was the reason for the famous eclipse expedition of 1919, led by Arthur Eddington, which found a result consistent with Einstein's prediction (figure 14). The third test was the prediction that clocks run slow in a gravitational potential — that is, a clock closer to the ground runs slow with respect to a clock at the top of a tower. This effect has also been measured experimentally. These were never, however, very impressive tests — the effects were always very small and various different theories can give the same results.

The situation has now changed dramatically — in 1993, Russell Hulse and Joseph Taylor were awarded the Nobel prize for a most

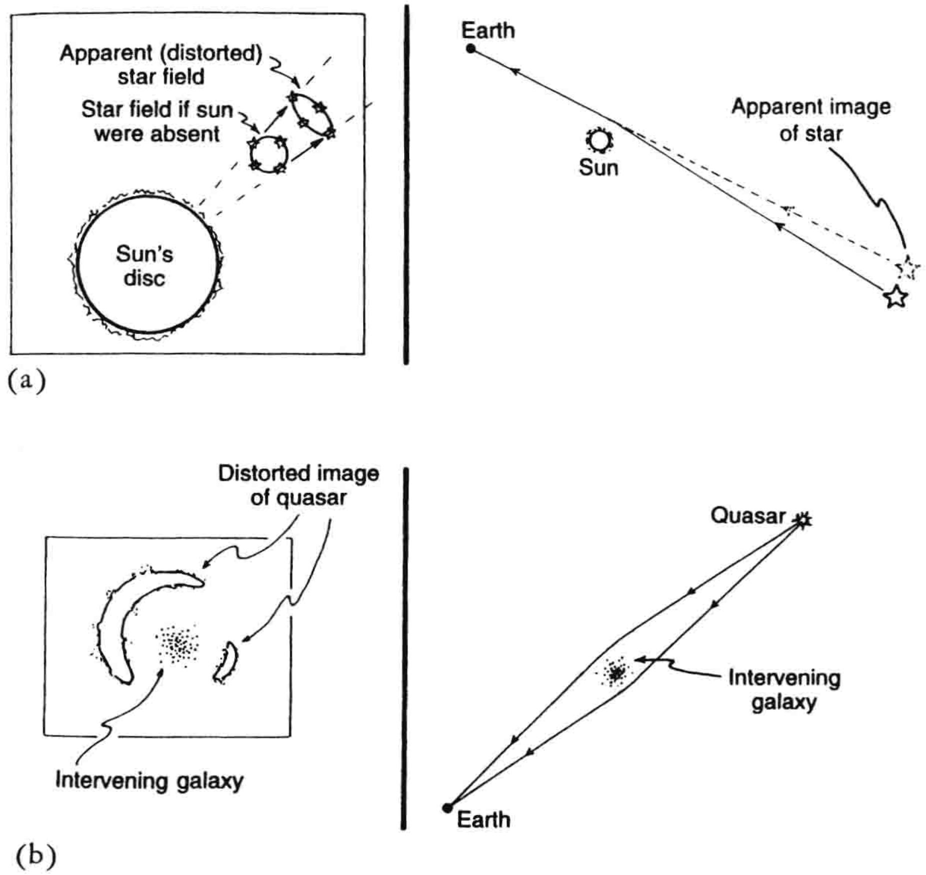
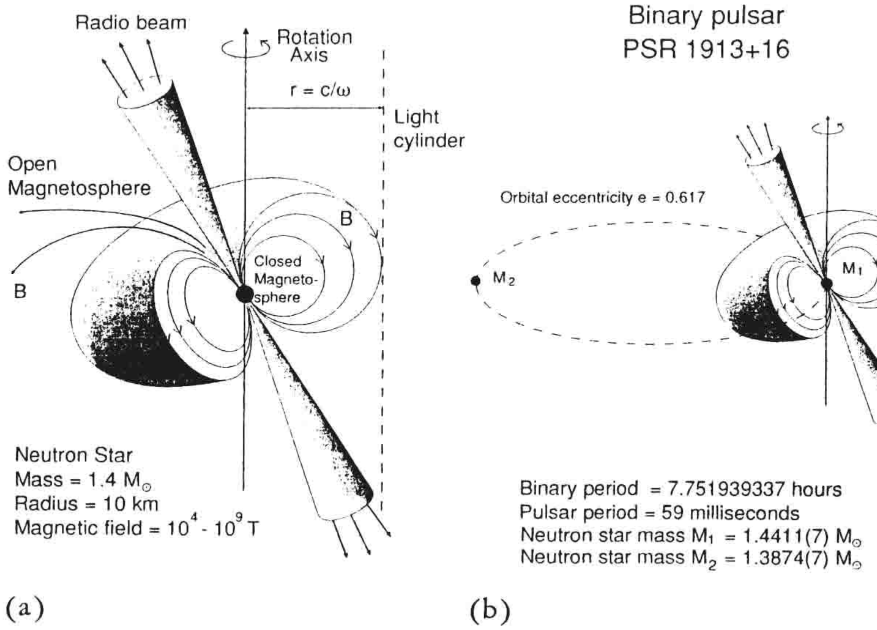

Apparent (distorted) star field
Star field if sun were absent
Sun's disc
(a)
Earth
Apparent image of star
Sun
Distorted image of quasar
Intervening galaxy
Quasar
Intervening galaxy
Earth
(b)

FIGURE 14. (a) A direct observational effect of light-cone tilting. The *Weyl* space-time curvature manifests itself as a distortion of the distant star field, here owing to the light-bending effect of the Sun's gravitational field. A circular pattern of stars would get distorted into an elliptical one. (b) Einstein's light-bending effect is now an important tool in observational astronomy. The mass of the intervening galaxy may be estimated by how much it distorts the image of a distant quasar.

extraordinary series of observations. Figure 15 shows the binary pulsar known as PSR 1913+16 — it consists of a pair of neutron stars, each of which is an enormously dense star that has mass about that of the Sun but is only a few kilometres in diameter. The neutron stars orbit around their common centre of gravity in highly elliptical orbits. One of them has a very strong magnetic field and particles get swung round and emit intense radiation that travels to the Earth, some 30,000 light years away, where it is observed as a series of well-defined pulses. All sorts of very precise observations have been made of the arrival times of these pulses. In particular, all the properties of the orbits of the two neutron stars can be worked out as well as all the tiny corrections due to General Relativity.

There is, in addition, a feature that is completely unique to general relativity, and not present at all in the Newtonian theory of gravity. That is that objects in orbit about each other radiate away energy in the form of gravitational waves. These are like light waves but are ripples in space-time rather than ripples in the electromagnetic field. These waves take energy away from the system at a rate that can be precisely calculated according to Einstein's theory, and the rate of loss of energy of the binary neutron star system agrees very precisely with the observations, as illustrated by figure 15(c), which shows the speed-up of the orbital period of the neutron stars, measured over twenty years of observation. These signals can be timed so precisely that, over twenty years, the accuracy with which the theory is known to be correct amounts to about one part in 10^{14}. This makes General Relativity the most accurately tested theory known to science.

There is a moral in this story — Einstein's motivations for devoting eight or more years of his life to deriving the General Theory were not observational or experimental. Sometimes people argue, "Well, physicists look for patterns in their experimental results and then they find some nice theory that agrees with these. Maybe this explains why mathematics and physics work so well


Radio beam
Rotation Axis
$r = c/\omega$
Light cylinder
Open Magnetosphere
B
B
Closed Magneto-sphere
Neutron Star
Mass = 1.4 $M_\odot$
Radius = 10 km
Magnetic field = 10^4 - 10^9 T
(a)
Binary pulsar
PSR 1913+16
Orbital eccentricity e = 0.617
M_2
M_1
Binary period = 7.751939337 hours
Pulsar period = 59 milliseconds
Neutron star mass M_1 = 1.4411(7) $M_\odot$
Neutron star mass M_2 = 1.3874(7) $M_\odot$
(b)

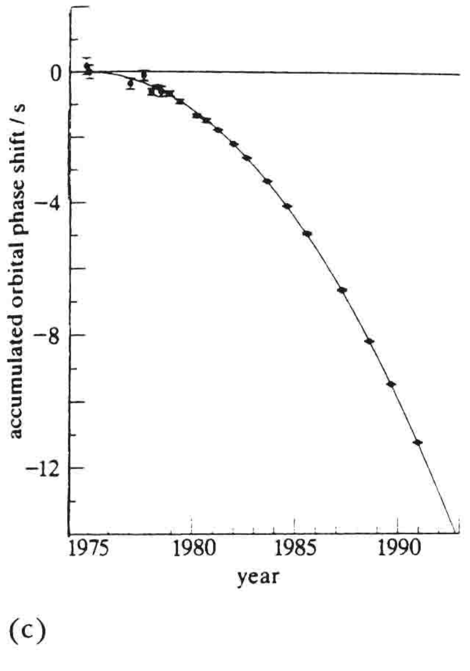


FIGURE 15. (a) A schematic diagram illustrating the properties of those neutron stars that are radio pulsars. Radio emission is emitted along the poles of the magnetic dipole that is misaligned with respect to the rotation axis of the neutron star. Sharply defined pulses are observed when the narrow beam of radiation is swept across the line of sight to the observer. (b) A schematic representation of the binary pulsar PSR 1913+16. The properties of the two neutron stars have been derived from very precise timing of the arrival times of the pulses using effects that are only present in Einstein's General Relativity. (c) The change of phase of the arrival times of the pulses from the binary pulsar PSR 1913+16, compared with the expected change due to the emission of gravitational radiation by the binary neutron star system.

together." But, in this case, things were not like that at all. The theory was developed originally without any observational motivation — the mathematical theory is very elegant and it is physically very well motivated. The point is that the mathematical structure is just there in nature, the theory really is out there in space — it has not been imposed upon nature by anyone. That is one of the essential points of this lecture. Einstein revealed something that was there. Moreover, it was not just some minor piece of physics he discovered — it is the most fundamental thing that we have in nature, the nature of space and time.

Here is a very clear case — it goes back to my original diagram concerning the relation between the world of mathematics and the physical world (figure 3). In General Relativity, we have some kind of structure that really does underlie the behaviour of the physical world in an extraordinarily precise way. The way in which these fundamental features of our world are discovered is often not by looking at the way in which nature behaves, although that is obviously very important. One has to be prepared to throw out theories that might appeal for all sorts of other reasons but that do not fit the facts. But here we have a theory that does fit the facts with extraordinary accuracy. The accuracy involved is about twice as many figures as one has in Newtonian theory; in other words, General Relativity is known to be correct to one part in 10^{14} whereas Newtonian theory was tested only to one part in 10^7. The improvement is similar to the increase in accuracy with which Newton's theory was known to be correct between the seventeenth century and now. Newton knew his theory was correct to about one part in 1,000, whereas now it is known to be accurate to one part in 10^7.

Einstein's General Relativity is just a theory, of course. What about the structure of the actual world? I said this lecture would not be botanical but, if I talk about the Universe as a whole, that is not being botanical, since I will consider only the one Universe as a whole that is given to us. There are three types of standard model that come out of Einstein's theory and these are defined by one parameter, which is, in effect, the one denoted by k in figure 16. There is another parameter that sometimes appears in cosmological arguments that is known as the cosmological constant. Einstein regarded his introduction of the cosmological constant into his equations of General Relativity as his greatest mistake and so I shall leave it out too. If we are forced to bring it back, well, we shall have to live with it.

Assuming the cosmological constant is zero, the three types of Universe, which are described by the constant k, are illustrated in figure 16. In the diagram, k takes values 1, 0, and -1, because all the other properties of the models have been scaled away. A better way would have been to talk about the age or scale of the Universe, and then one would have a continuous parameter but, qualitatively, the three different models can be thought of as being defined by the curvature of the space sections of the Universe. If the space sections of Universe are flat, they have zero curvature and $k = 0$. If the space sections are positively curved, meaning that the Universe closes in on itself, then $k = +1$. In this case the Universe has an initial singular state, the Big Bang, which marks the beginning of the Universe. It expands to a maximum size and then recollapses to a Big Crunch. Alternatively, there is the $k = -1$ case, in which the Universe expands forever. The $k = 0$ case is the limiting boundary between the $k = 1$ and $k = -1$ cases. Beside the diagrams, I have shown the radius-time relations for these three types of Universe in figure 16(d). The radius can be thought of as some typical scale in the Universe and it can be seen that only the case $k = +1$ collapses to a Big Crunch, while the other two expand indefinitely.

I want to consider the $k = -1$ case in a little more detail — it is perhaps the most difficult of the three to come to terms with. There are two reasons for being interested in this case particularly. One reason is that, if you take the observations as they exist at the moment at their face value, it is the preferred model. According

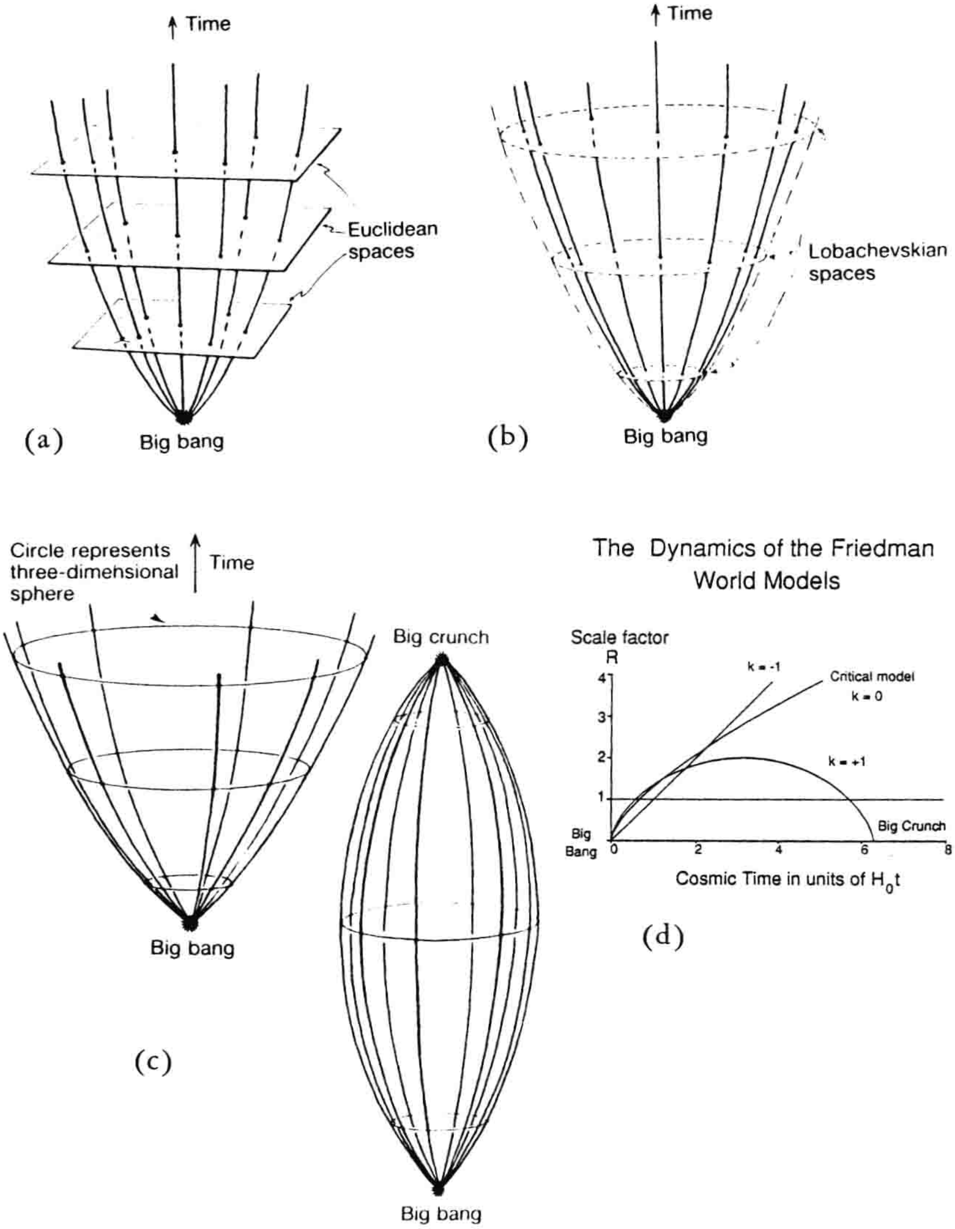

Time
Euclidean spaces
(a)
Big bang
Time
Lobachevskian spaces
(b)
Big bang
Circle represents three-dimensional sphere
Time
Big bang
(c)
Big crunch
Big bang
The Dynamics of the Friedman World Models
Scale factor R
4
3
2
1
k = -1
Critical model k = 0
k = +1
Big Crunch
Big Bang
0
2
4
6
8
Cosmic Time in units of H_0t
(d)

FIGURE 16. (a) Space-time picture of an expanding universe with Euclidean spatial sections (two space dimensions depicted), $k=0$. (b) As in (a) but for an expanding universe with Lobachevskian spatial sections, $k=-1$. (c) As in (a), but for an expanding universe with spherical spatial sections $k=+1$. (d) The dynamics of the three different types of Friedman model.

to General Relativity, the curvature of space is determined by the amount of matter present in the Universe and there doesn't seem to be enough to close the geometry of the Universe. Now, it may be that there is a lot of dark or hidden matter, which we do not yet know about. In this case, the Universe could be one of the other models but, if there is not a lot of extra matter, much more than we believe must be present within the optical images of galaxies, then the Universe would have $k = -1$. The other reason is that it is the one I like the best! The properties of $k = -1$ geometries are particularly elegant.

What do the $k = -1$ universes look like? Their spatial sections have what is known as hyperbolic or Lobachevski geometry. To get a picture of a Lobachevski geometry it is best to look at one of Escher's prints. He made a number of prints that he called *Circle Limits*, and *Circle Limit 4* is shown in figure 17. This is Escher's description of the Universe — you see it is full of angels and devils!

A point to note is that it looks as though the picture gets very crowded towards the edge of the limit circle. This occurs because this representation of hyperbolic space is drawn on an ordinary plane sheet of paper, in other words, in Euclidean space. What you have to imagine is that all the devils are supposed to be actually exactly the same size and shape so that, if you happened to live in this Universe towards the edge of the diagram, they would look exactly the same to you as the ones in the middle of the diagram. This picture gives some impression of what is going on in Lobachevski geometry — as you walk from the centre out to the edge, you have to imagine that, because of the way the picture of the geometry has had to be distorted, the actual geometry there is exactly the same as it is in the middle, so that the geometry all about you remains the same no matter how you move.

This is perhaps the most surprising example of a well-defined geometry. But Euclidean geometry is, in its way, just as remarkable. Euclidean geometry provides a wonderful illustration of the

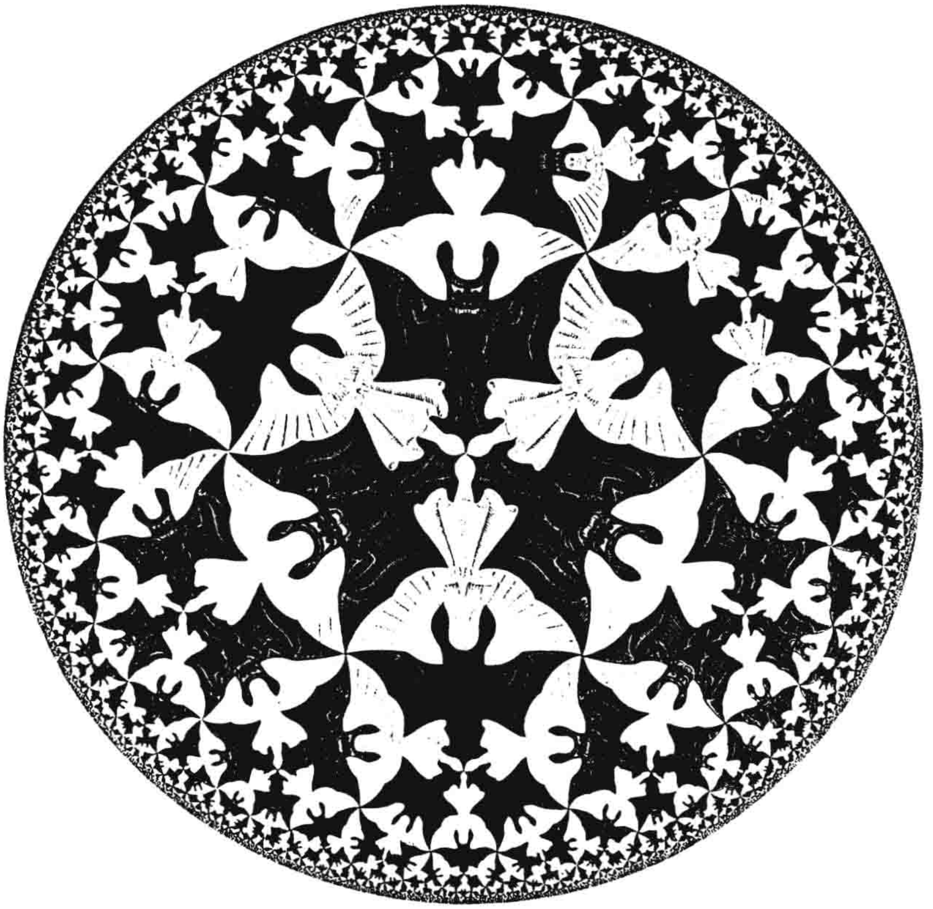


FIGURE 17. *Circle Limit 4* by M. C. Escher.

relation between mathematics and physics. The geometry is a part of mathematics, but the Greeks also thought of it as a description of the way the world is. Indeed, it turns out to be an extraordinarily accurate description of the way the world actually is — not utterly accurate because Einstein's theory tells us that space-time is slightly curved in various ways, but it is an extraordinarily accurate description of the world nevertheless. People used to worry about whether or not other geometries were possible. In particular, they worried about what is known as *Euclid's fifth postulate*. This can be reformulated as the statement that, if there is a line in a plane and there is a point outside that line, then there is a unique parallel to this line through that point. People used to think that maybe this could be proved from the other more obvious axioms of Euclidean geometry. It turns out that it is not possible, and from this the notion of non-Euclidean geometry arose.

In non-Euclidean geometries, the angles of a triangle do not add up to 180°. This is another example where you think that things must become more complicated because, in Euclidean geometry, the angles of a triangle do add up to 180°. But then, in the non-Euclidean geometry, if you take the sum of the angles of a triangle away from 180°, you find that this difference is proportional to the area of the triangle. In Euclidean geometry, the area of a triangle is a complicated thing to write down in terms of angles and lengths. In non-Euclidean, Lobachevskian, geometry, there is this wonderfully simple formula, due to Johann Heinrich Lambert, which enables the area of the triangle to be found. In fact, Lambert derived his formula before non-Euclidean geometry was discovered and I have never quite understood that!

There is another very important point here, which concerns the real numbers. These are absolutely fundamental to Euclidean geometry. They were essentially introduced by Eudoxus in the fourth century B.C. and they are still with us. They are the numbers that describe all our physics. As we shall see later, complex numbers are needed too, but they are based upon real numbers.

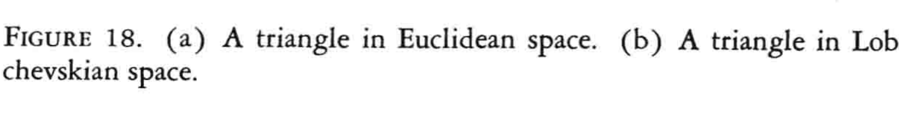

(a)
γ
α
β
$\alpha+\beta+\gamma=180°$
(b)
γ
α
β
$180°-\alpha-\beta-\gamma=$const.×area

FIGURE 18. (a) A triangle in Euclidean space. (b) A triangle in Lobachevskian space.

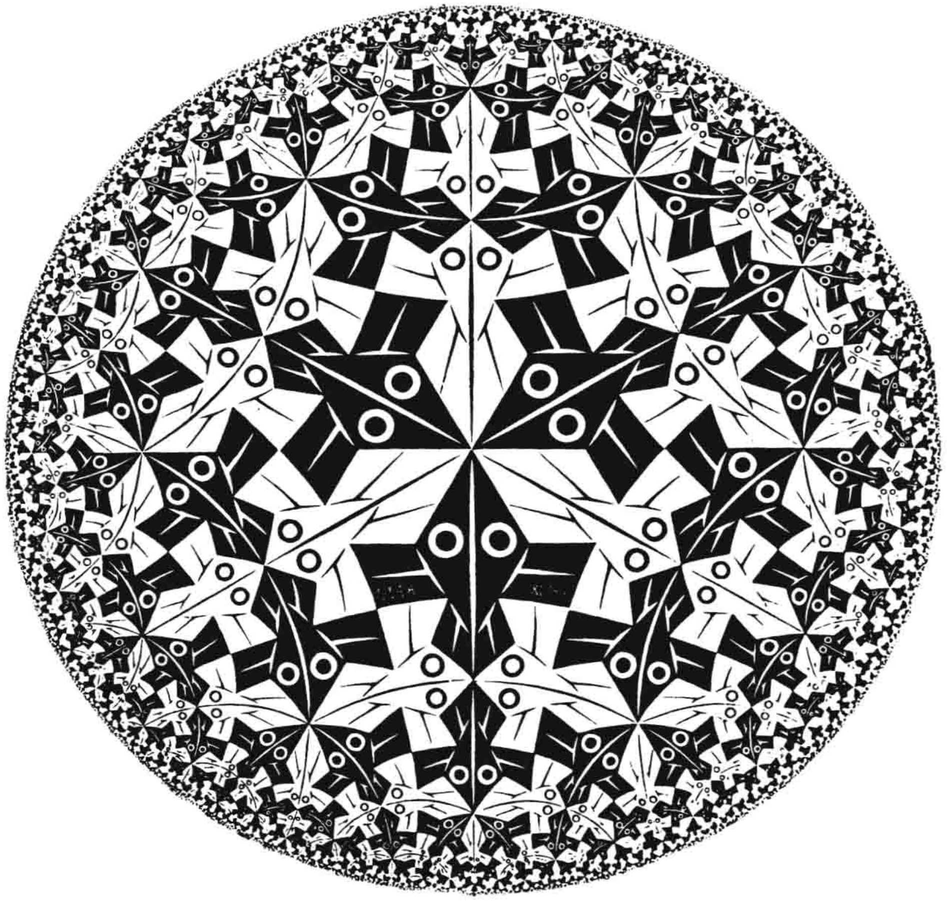


FIGURE 19. *Circle Limit 1* by M. C. Escher.

Let us look at another of the Escher prints to see how the Lobachevski geometry works. Figure 19 is even nicer than figure 18 for understanding this geometry because the "straight lines" are more obvious. They are represented by arcs of circles that meet the boundary at right angles. So, if you were a Lobachevskian person, and lived in this geometry, what you would think of as a straight line would be one of these arcs. You can see these clearly in figure 19 — some of them are straight lines through the centre of the diagram, but all the others are curved arcs. Some of these "straight lines" are shown in figure 20. In that diagram, I have indicated a point that is not on the straight line crossing the diagram. Lobachevskian people can draw two (and more) separate lines parallel to the diagonal through that point, as I have indicated. Thus, the parallel postulate is violated in this geometry. Furthermore, you can draw triangles and work out the sums of angles of the triangles in order to work out their areas. This may give you some taste for the nature of hyperbolic geometry.

Let me give another example. I said that I like hyperbolic, Lobachevskian geometry the best. One of the reasons is that its group of symmetries is exactly the same as the one that we have already encountered, namely, the Lorentz group — the group of Special Relativity, or the symmetry group of the light-cones of relativity. To see that it is, I have drawn a light-cone in figure 21 but with some extra bits drawn on. I have had to suppress one of the space dimensions in order to depict it in three-dimensional space.

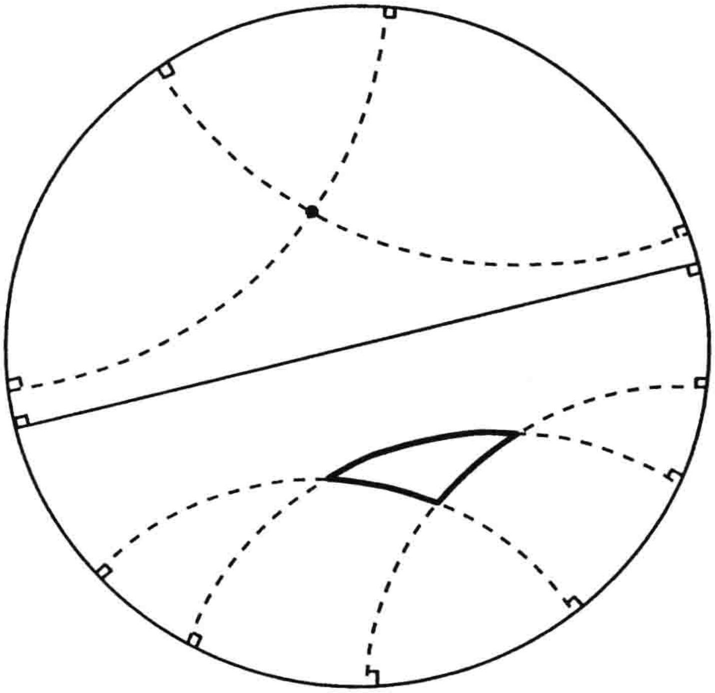


FIGURE 20. Illustrating the geometry of Lobachevskian (hyperbolic) space.

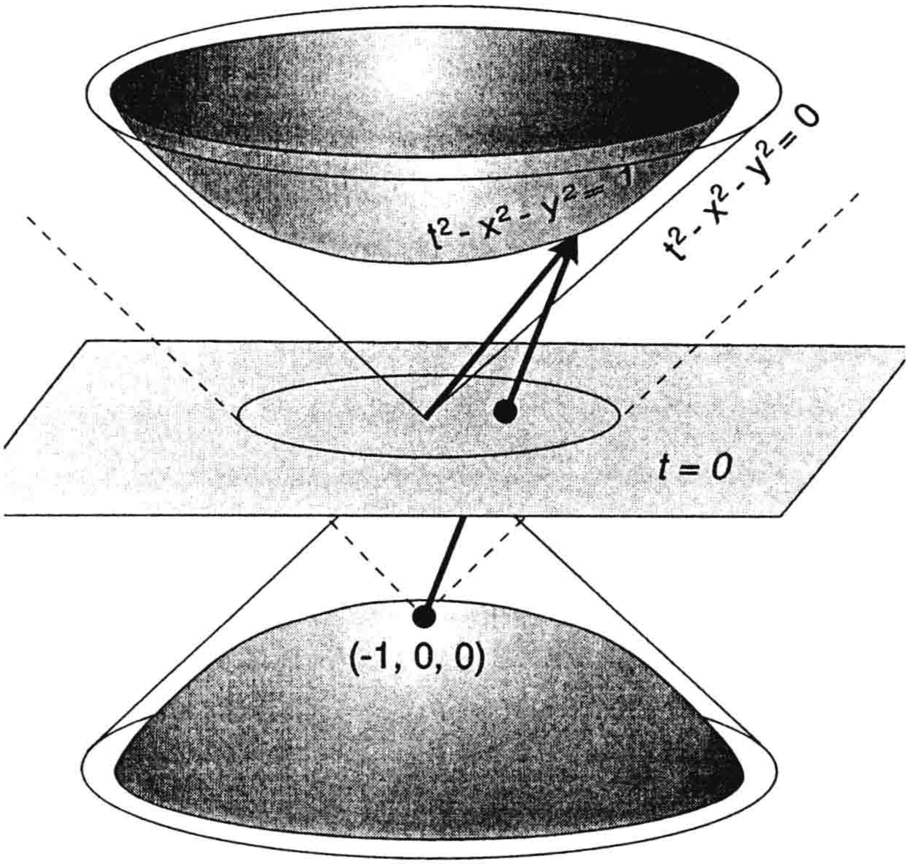


FIGURE 21. Illustrating Lobachevskian space-time geometry. The circle drawn on the plane $t = 0$ is the Poincaré disc.

The light-cone is described by the usual equation shown on the diagram:

$$t^2 - x^2 - y^2 = 0$$

The bowl-shaped surfaces shown above and below are located at "unit distance" from the origin in this Minkowskian geometry. ("Distance" in Minkowskian geometry is actually *time* — the proper time that is physically measured by moving clocks.) Thus, these surfaces represent the surface of a "sphere" for the Minkowskian geometry. It turns out that the intrinsic geometry of the "sphere" is actually Lobachevskian (hyperbolic) geometry. If you consider an ordinary sphere in Euclidean space, you can rotate it around, and the group of symmetries is that of rotating the sphere around. In the geometry of figure 21, the group of symmetries is the group of symmetries associated with the surface shown in the diagram, in other words, with the Lorentz group of rotations. This symmetry group describes how space and time transform when a particular point in space-time is fixed — rotating the space-time about in different ways. We now see, with this representation, that the group of symmetries of Lobachevskian space is essentially just the same as the Lorentz group.

We can carry out a Minkowskian version of the stereographic projection, as illustrated in figure 10(b). The equivalent of the south pole is now the point at $(-1, 0, 0)$ and we project points from the upper bowl-shaped surface to the flat surface at $t = 0$, which is the equivalent of the equatorial plane in figure 10(b). In this procedure, we project all the points on the upper surface to the plane at $t = 0$. The projected points all lie inside a disc in the plane at $t = 0$, and this disc is sometimes referred to as the Poincaré disc. This is precisely how Escher's circle limit diagrams come about — the entire hyperbolic (Lobachevskian) surface has been mapped onto the Poincaré disc. Furthermore, this mapping does all the things that the projection of figure 10(b) does — it preserves angles and circles and it all comes out geometrically in a very nice way. Well, perhaps I am getting carried away here by my enthusiasm — I am afraid that is what mathematicians do when they get stuck into something.

The intriguing point is that, when you get carried away by something like the geometry of the above problem, the analysis and the results have an elegance that sustains them, while analyses that do not possess this mathematical elegance peter out. There is something particularly elegant about hyperbolic geometry. It would be awfully nice, at least to the likes of me, if the Universe were built that way too. Let me say that I have various other reasons for believing this. Many other people do not like these open, hyperbolic universes — they frequently prefer closed universes, such as those illustrated in figure 16(c), which are nice and cosy. Well, actually, the closed universes are pretty big still. Alternatively, many people like flat world models as in figure 16(a) because there is a certain type of theory of the early Universe, the *inflationary theory*, which suggests that the geometry of the Universe should be flat. I should say that I do not really believe these theories.

These three standard types of model of the Universe are what are known as the Friedman models and they are characterised by the fact that they are very, very symmetrical. They are initially expanding models but at any moment the Universe is perfectly uniform everywhere. This assumption is built into the structure of the Friedman models and it is known as the *cosmological principle*. Wherever you are, the Friedman universe looks the same in all directions. It turns out that our actual Universe is like this to a remarkable degree. If Einstein's equations are right, and I have shown that the theory agrees with observation to a quite remarkable degree, then we are led to take the Friedman models seriously. All these models have this awkward feature, known as the Big Bang, where everything goes wrong, right at the beginning. The Universe is infinitely dense, infinitely hot, and so on — something has gone badly wrong with the theory. Nonetheless, if you accept that this very hot, dense phase took place, you can make predictions about what the thermal content of the Universe should be today and one of these expectations is that there should be a uniform background of black-body radiation all about us at the present day. Precisely this type of radiation was discovered by Arno Penzias and Robert Wilson in 1965. The most recent observations of the spectrum of this radiation, which is known as the Cosmic Microwave Background Radiation, by the COBE satellite show that it has precisely a black body spectrum with quite extraordinary precision.

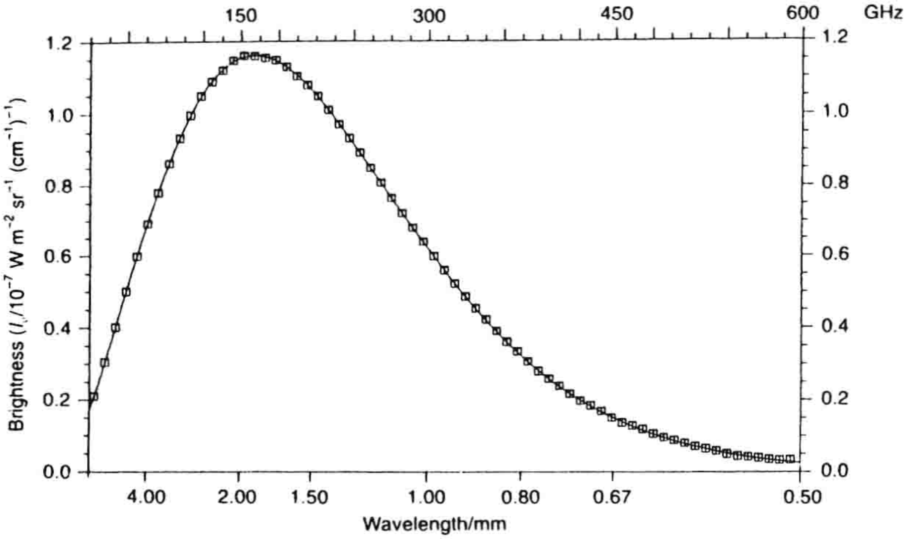


FIGURE 22. The precise agreement between the COBE measurements of the spectrum of the cosmic microwave background radiation and the expected "thermal" nature of the Big Bang's radiation.

All cosmologists interpret the existence of this radiation as evidence that our Universe went through a hot, dense phase. This radiation is thus telling us something about the nature of the early Universe — it is not telling us everything, but something like the Big Bang did take place. In other words, the Universe must have been very like the models illustrated in figure 16.

There is one other very important discovery made by the COBE satellite and this is that, although the Cosmic Microwave Background Radiation is remarkably uniform and its properties can all be accounted for very beautifully mathematically, the Universe is not quite perfectly uniform. There are tiny but real irregularities in the distribution of the radiation over the sky. In fact, we expect that these tiny irregularities must be present in the early Universe — we are here to observe the Universe and we are certainly not just a uniform smudge. The Universe is probably more like the pictures illustrated in figure 23. To show how open-minded I am, I am using as examples both an open and a closed Universe.

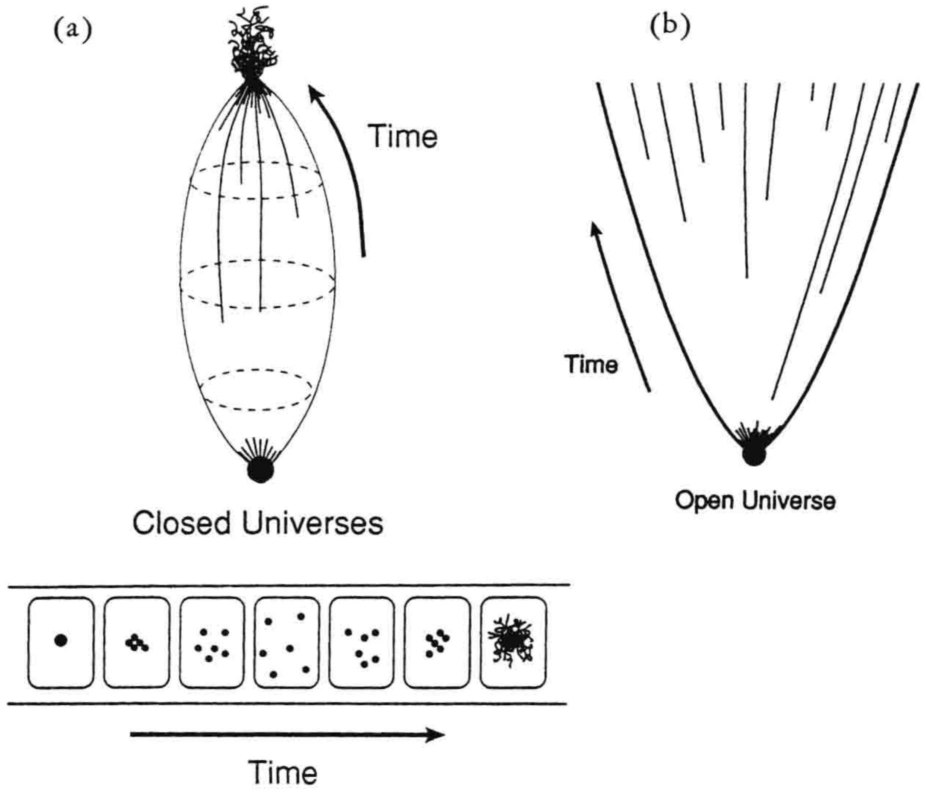


FIGURE 23. (a) The evolution of a closed world model with the formation of black holes as objects of various types reach the end points of their evolution. It can be seen that there is expected to be a horrible mess at the Big Crunch. The sequence of events is also shown as a "film-strip." (b) The evolution of the open model with the formation of black holes.

In the closed Universe, the irregularities will develop to form real observable structures — stars, galaxies, and the like — and, after a while, black holes will form, through the collapse of stars, through the accumulation of mass at the centres of galaxies, and so on. These black holes all have singular centres, much like the Big Bang in reverse. However, it is not as simple as that. According to the picture we have developed, the initial Big Bang is a nice, symmetrical, uniform state but the end point of the closed model is a horrible mess — all the black holes finally coming together and producing an incredible jumble at the final Big Crunch, as in figure 23(a). The evolution of this closed model is illustrated schematically by the film strip shown in figure 23(b). In the case of an open universe model, the black holes are still formed as well; there is still an initial singularity and singularities formed at the centres of the black holes.

I emphasise these features of the standard Friedman models to show that there is a great difference between what we seem to see in the initial state and what we expect to find in the remote future. This problem is connected with the fundamental law of physics known as the Second Law of Thermodynamics.

We can understand this law in simple everyday terms. Imagine a glass of wine perched on the edge of the table. It might fall off the table, smash to pieces, and the wine spill all over the carpet (figure 24). There is nothing in Newtonian physics that tells us that the reverse process cannot happen. However, that is never observed — you never see wine glasses reassembling themselves and the wine being sucked up out of the carpet and into the reassembled glass. So far as the detailed laws of physics are concerned, the one direction of time is just as good as the other. To understand this difference, we need the Second Law of Thermo-

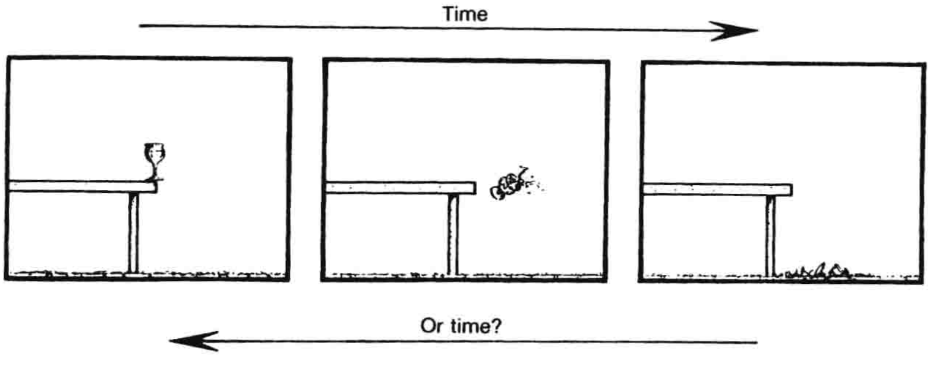


FIGURE 24. The laws of mechanics are time-reversible; yet the time-ordering of such a scene from the right frame to the left is something that is never experienced, whereas that from the left frame to the right would be commonplace.

dynamics, which tells us that the entropy of the system increases with time. This quantity called entropy is lower when the glass is on the table as compared with when it is shattered on the floor. According to the Second Law of Thermodynamics, the entropy of the system has increased. Roughly speaking, entropy is a measure of the disorder of a system. To express this concept more precisely, we have to introduce the concept of a *phase space*.

A phase space is a space of an enormous number of dimensions and each point of this multidimensional space describes the positions and momenta of all the particles that make up the system under consideration. In figure 25, we have selected a particular point in this huge phase space that represents where all the particles are located and how they are moving. As the system of particles evolves, the point moves to somewhere else in the phase space and I have shown it wiggling about from one point in phase space to another.

This wiggly line represents the ordinary evolution of the system of particles. There is no entropy there yet. To get entropy, we

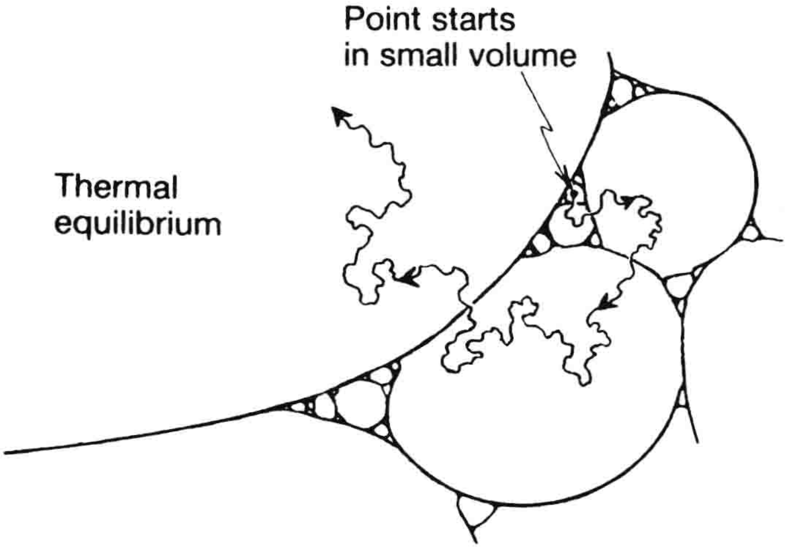


FIGURE 25. The Second Law of Thermodynamics in action: as time evolves, the phase-space point enters compartments of larger and larger volumes. Consequently, the entropy continually increases.

have to draw little bubbles around regions by lumping together those different states that you cannot tell apart. That may seem a bit obscure — what do you mean by "Cannot tell apart"? Surely that depends upon those who are looking and how carefully they look? Well, it is one of the slightly tricky questions of theoretical physics to say exactly what you mean by entropy. Essentially, what is meant is that you have to group states together according to what is known as "coarse-graining," that is, according to those things that you cannot tell apart. You take all those that, say, lie in this region here and lump them together, you look at the volume of that region, take the logarithm of the volume, and multiply it by the constant known as Boltzmann's constant, and that is the entropy. What the Second Law of Thermodynamics tells us is that the entropy increases. What it is telling you is actually something rather silly — all it is saying is that, if the system starts off in a little tiny box and it is allowed to evolve, it moves into bigger and bigger boxes. It is very likely that this happens because, if you look at the problem carefully, the bigger boxes are absolutely stupendously huger than the neighbouring little boxes. So, if you find yourself in one of the big boxes, there is absolutely no chance of getting back into a smaller box. That is all there is to it. The system just wanders about in phase space getting into bigger and bigger boxes. That is what the Second Law is telling us. Or is it?

Actually, that is only half the explanation. It tells us that, if we know the state of the system now, we can tell the most likely state in the future. But it tells us the completely wrong answer if we try to use the same argument backwards. Suppose the glass is sitting on the edge of the table. We can ask, "What is the most likely way by which it could have got there?" If you use the argument we have just given backwards, you would conclude that the most likely thing is that it started as a great mess on the carpet and then picked itself up off the carpet and reassembled itself on the table. This is clearly not the correct explanation — the correct explanation is that someone put it there. And that person put it there for some reason, which was in turn due to some other reason, and so on. The chain of reasoning goes back and back to lower and lower entropy states in the past. The correct physical curve is illustrated in figure 26 — the entropy goes down and down and down in the past.

The reason why the entropy increases in the future is explained by moving into larger and larger boxes — why it goes down in the past is something completely different. There must have been something that pulled it down in the past. What pulled it down in the past? As we go into the past, the entropy gets smaller and smaller until eventually we end up at the Big Bang.

There must have been something very, very special about the Big Bang, but exactly what that was is a controversial issue. One popular theory, which I said I did not believe but which a lot of people are very keen on, is the idea of the inflationary universe. The idea is that the Universe is so uniform on the large scale because of something that was supposed to have taken place in the

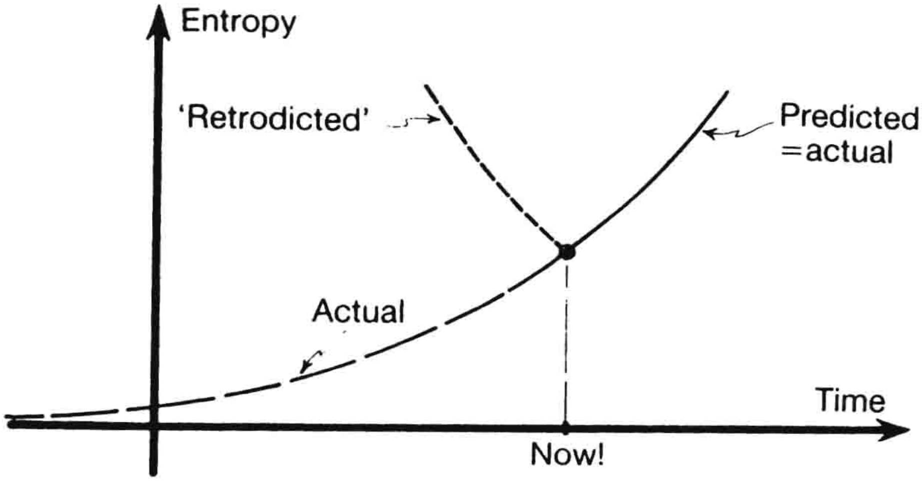


FIGURE 26. If we use the argument depicted in figure 25 in the reverse direction in time, we "retrodict" that the entropy should also increase in the *past*, from its value now. This is in gross contradiction with observation.

very earliest phases of the expansion of the Universe. It is supposed that an absolutely enormous expansion took place when the Universe was only about 10^{-36} seconds old and the idea is that, no matter what the Universe looked like in these very early stages, if you expand it by a huge factor of about 10^{60}, then it will look flat. In fact, this is one reason why these people like the flat Universe.

But, as it stands, the argument does not do what it is supposed to do — what you would expect in this initial state, if it were randomly chosen, would be a horrendous mess and, if you expand that mess by this huge factor, it still remains a complete mess. In fact, it looks worse and worse the more it expands (figure 27).

So the argument by itself does not explain why the Universe is so uniform. We need a theory that tells you what the Big Bang was really like. We do not know what that theory is, but we know that it has to involve a combination of large-scale and small-scale physics. It has to involve quantum physics as well as classical physics. Furthermore, I would claim that the theory must also have as one of its implications that the Big Bang was as uniform as we observe it to be. Maybe such a theory will end up producing a hyperbolic, Lobachevskian universe, like the picture I prefer, but I shall not insist upon that.

Let us return to the pictures of the closed and open Universes again (figure 28). In addition, I have included a picture of the formation of a black hole, which will be well known to the ex-

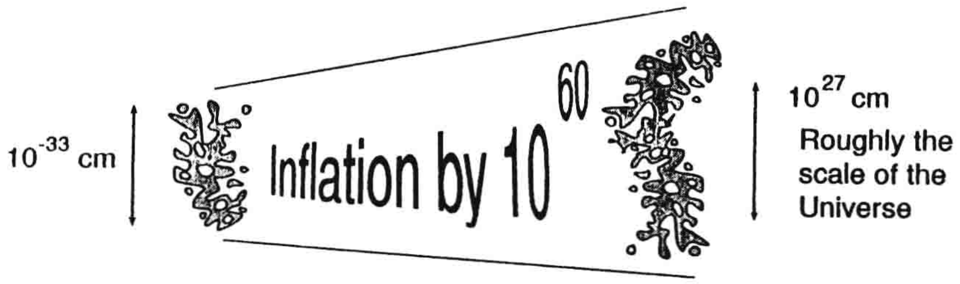


FIGURE 27. Illustrating the problem of the inflation of irregularities in the early Universe.

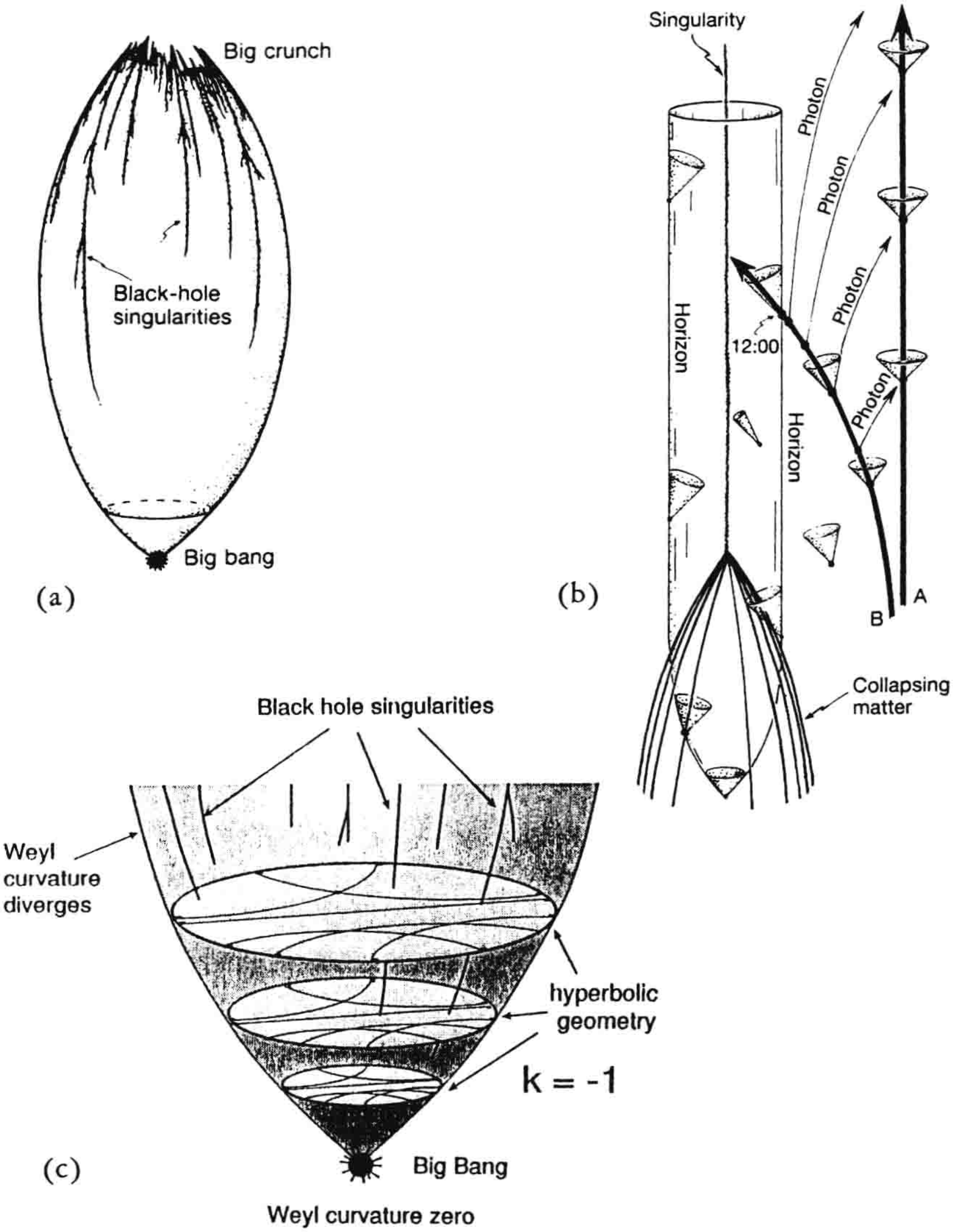

Big crunch
Black-hole singularities
Big bang
(a)
Singularity
Photon
Photon
Photon
Photon
Horizon
12:00
Horizon
(b)
B
A
Collapsing matter
Black hole singularities
Weyl curvature diverges
hyperbolic geometry
k = -1
(c)
Big Bang
Weyl curvature zero

FIGURE 28. (a) The entire history of a closed universe that starts from a uniform low-entropy Big Bang with $Weyl = 0$ and ends with a high-entropy Big Crunch — representing the congealing of many black holes — with $Weyl \to \infty$. (b) A space-time diagram depicting collapse to a black hole. (c) The history of an open universe, again starting from a uniform low-entropy Big Bang with $Weyl = 0$.

perts. Matter collapsing into a black hole produces a singularity and that is what the dark lines on the space-time diagrams of the Universe represent. I want to introduce a hypothesis that I call the *Weyl curvature hypothesis.* This is not an implication of any known theory. As I have said, we do not know what the theory is, because we do not know how to combine the physics of the very large and the very small. When we do discover that theory, it should have as one of its consequences this feature that I have called the Weyl curvature hypothesis. Remember that the Weyl curvature is that bit of the Riemann tensor that causes distortions and tidal effects. For some reason we do not yet understand, in the neighbourhood of the Big Bang, the appropriate combination of theories must result in the Weyl tensor being essentially zero, or rather be constrained to be very small indeed.

That would give us a Universe like that shown in figure 28(c) and not like that in figure 28(a). The Weyl curvature hypothesis is time-asymmetrical and it applies only to the past type singularities and not to the future singularities. If the same rule of vanishing Weyl tensor that I have applied in the past also applied to the future of the Universe, in the closed model, you would end up with a dreadful looking Universe with as much mess in the past as in the future (figure 29). This looks nothing like the Universe we live in.

What is the probability that, purely by *chance*, the Universe had an initial singularity looking even remotely as it does? The probability is less than one part in $10^{10^{123}}$. Where does this estimate come from? It is derived from a formula by Jacob Beckenstein and Stephen Hawking concerning black hole entropy and, if you apply it in this particular context, you obtain this enormous answer. It depends how big the Universe is and, if you adopt my own favorite Universe, the number is, in fact, infinite.

What does this say about the precision that must be involved in setting up the Big Bang? It is really very, very extraordinary. I have illustrated the probability in this cartoon of the Creator,

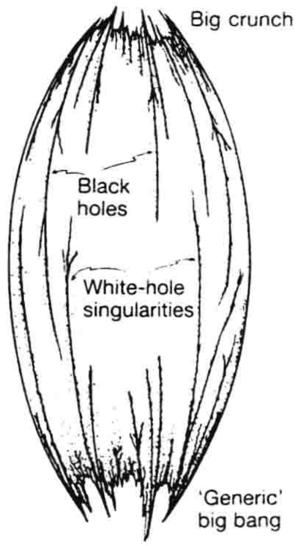


FIGURE 29. If the constraint *Weyl* = 0 is removed, then we have a high-entropy Big Bang also, with *Weyl* → ∞ there. Such a universe would be riddled with white holes, and there would be no Second Law of Thermodynamics, in gross contradiction with experience.

finding a very tiny point in that phase space that represents the initial conditions from which our Universe must have evolved if it is to resemble remotely the one we live in (figure 30). To find it, the Creator has to locate that point in phase space to an accuracy of one part in $10^{10^{123}}$. If I were to put one zero on each elementary particle in the Universe, I still could not write the number down in full. It is a stupendous number.

I have been talking about precision — how mathematics and physics fit together with extraordinary accuracy. I have also talked about the Second Law of Thermodynamics, which is often thought of as a rather floppy law — it concerns randomness and chance, and yet there is something very precise hiding underneath this law. As applied to the Universe, it has to do with the precision with which the initial state was set up. This precision must be something to do with the union of quantum theory and general relativity, a theory we do not have.

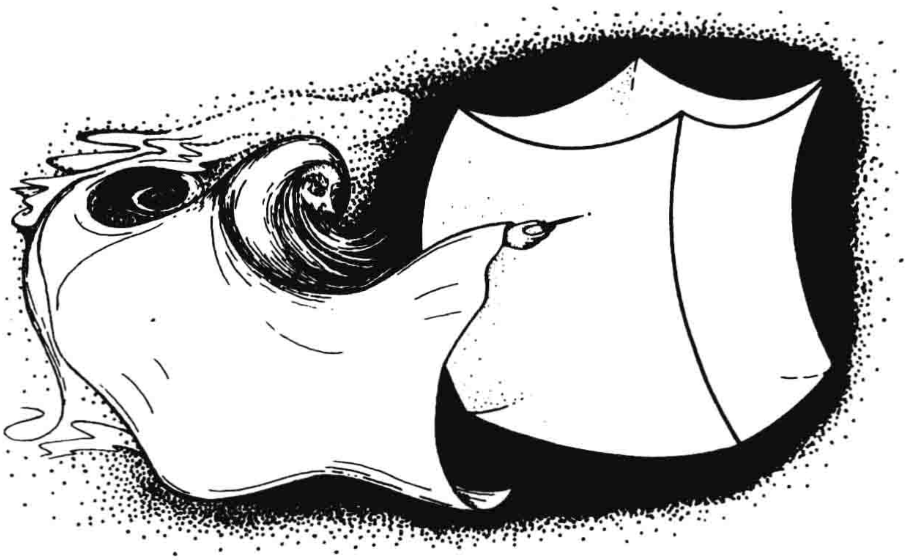


FIGURE 30. In order to produce a universe resembling the one in which we live, the Creator would have to aim for an absurdly tiny volume of the phase space of possible universes — about $1/10^{10^{123}}$ of the entire volume, for the situation under consideration. (The pin and the spot aimed for are not drawn to scale!)

249

NEW IDEAS FOR THE GOOGLY GRAVITON

Twistor Newsletter **40** (1996)

This represents the beginning of some new thoughts about the 'googly' problem for gravitation, which appear to have gradually opened up the subject again. See Chapters 263, 272, 289, and 294.

New Ideas for the Googly Graviton

In my articles in $\mathbb{T}\mathbb{N}$ 31 (pp. 6–8), 32 (1–5), 38 (1–9), 39 (1–5), in *Gravitation and Modern Cosmology* (eds. A. Zichichi, N. de Sabbata, & N. Sánchez; Plenum 1991), in *Twistor Theory* (ed. S. Huggett; Marcel Dekker 1995); and in L.J.M. & R.P. $\mathbb{T}\mathbb{N}$ 37 (1–6), J.F. $\mathbb{T}\mathbb{N}$ 37 (7–9), J-P.N. $\mathbb{T}\mathbb{N}$ 39 (6–10), and elsewhere, it has been suggested that the appropriate twistor-space procedure for encoding the structure of an arbitrary vacuum space-time $\mathcal{M}$ would be to find the space $\mathcal{T}$ of "charges" for helicity $\frac{3}{2}$ massless fields — or, rather, for potentials $\sigma_{A'B'}^{\ \ \ C}$ for such fields. This suggestion was based on the following two observations:

(1) the field equations (Dirac or Rarita-Schwinger) for $\sigma_{A'B'}^{\ \ \ C}$ are *consistent* iff $\mathcal{M}$ is Ricci-flat,

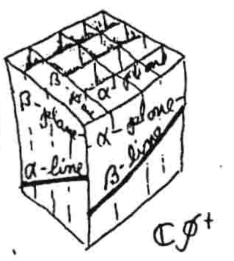
(2) in Minkowski space $\mathbb{M}$, the space of *charges* for $\sigma_{A'B'}^{\ \ \ C}$ is standard twistor space $\mathbb{T}$.

The programme, then, is to try to find the appropriate notion of "charge" for such fields for a general Ricci-flat $\mathcal{M}$, and then to try to see how to reconstruct $\mathcal{M}$ (with vacuum equations automatically satisfied) from the "twistor space" $\mathcal{T}$ of such charges.

In general, there appear to be severe difficulties standing in the way of constructing an appropriate twistor space in accordance with these ideas (cf. particularly R.P. in $\mathbb{T}\mathbb{N}$ 33 (1–6) and the above ref. in *Grav. & Mod. Cosmol.* 1991). The most plausible procedure for an asymptotically flat (analytic) $\mathcal{M}$ seems to be to go to infinity, say to $\mathscr{I}^+$, and see how best to define the space of helicity 3/2 charges in terms of "glitches" in the potentials at $\mathscr{I}^+$ or $\mathbb{C}\mathscr{I}^+$. These glitches would have to be not only in the first potential $\sigma_{A'B'}^{\ \ \ C}$ but also in a second potential $\rho_{A'}^{\ BC}$, all restricted to $\mathscr{I}^+$ (or $\mathbb{C}\mathscr{I}^+$). Recent work by Jorge Frauendiener, Ted Newman, and Joy Ghose has explored the nature of $\mathcal{T}$ constructed in this way, and seems to confirm my general expectation that $\mathcal{T}$ ought to turn out to be the *asymptotic twistor space* of $\mathcal{M}$.

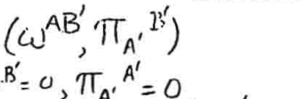
However, according to the understanding that has existed to date, the asymptotic twistor space of $\mathcal{M}$ encodes only the anti-self-dual (ASD) part of the radiation field of $\mathcal{M}$ — according to the $\mathcal{H}$-space/non-linear graviton construction — and not the self-dual (SD) part. (It should be borne in mind that all this information is to be encoded holomorphically, the operation of complex conjugation, or reality structure, not being something that one may call upon. Thus, we are, for the time being at least, operating within the framework of complex general relativity.) Accordingly, it appears to be essential to re-examine the googly programme for ascertaining how the SD part of the radiation field may also be holomorphically encoded in the structure of $\mathcal{T}$ (and not simply in that of the dual asymptotic twistor space $\mathcal{T}^*$). Ultimately, some information about sources for the gravitational field would also have to be encoded somehow. This article describes various ideas that, I believe, point to some significant new progress towards a solution of the googly problem.

Let us first recall the way in which asymptotic twistor space arises in terms of $\mathbb{C}\mathscr{I}^+$. We may regard $\mathbb{C}\mathscr{I}^+$ as a union of α-planes, or alternatively as a union of β-planes. There are null geodesics on $\mathbb{C}\mathscr{I}^+$ of three types: α-lines, β-lines, and generators, the generators being limiting cases of α-lines and also of β-lines. Each α-line lies on a unique β-plane and each β-line on a unique α-plane (all on $\mathbb{C}\mathscr{I}^+$), each generator lying on one of each. The "finite" part of projective asymptotic twistor space $\mathbb{P}\mathcal{T}^b$ is simply the space of α-lines (and the corresponding dual twistor space $\mathbb{P}\mathcal{T}^{*b}$ is the space of β-lines).



In the flat case of $\mathbb{M}$, $\mathbb{P}\mathcal{T}^b = \mathbb{P}\mathbb{T}^b$, which is the ordinary twistor space $\mathbb{P}\mathbb{T}$, but with the line $\mathbb{P}\mathrm{I}$ (representing infinity) removed. The entire space $\mathbb{P}\mathbb{T}$ arises if we include the α-planes on $\mathbb{C}\mathscr{I}^+$ in addition to the α-lines, these various α-planes providing the points of $\mathbb{P}\mathrm{I}$. This procedure could be adopted in the case of a general $\mathbb{C}\mathscr{I}^+$ also, but then it turns out that the resulting space $\mathbb{P}\mathcal{T}$ is not a smooth complex manifold, being singular along $\mathbb{P}\mathrm{I}$. In some appropriate sense, it is the nature of this singularity along $\mathbb{P}\mathrm{I}$ that contains all the googly gravitational information.

One standard way to proceed is to blow up the twistor space, so the line $\mathbb{P}\mathrm{I}$ becomes a quadric surface $\mathbb{P}\mathrm{II}$: In the flat case, this is achieved by $\frac{1}{z} \rightsquigarrow \frac{1}{z}z\rceil$,

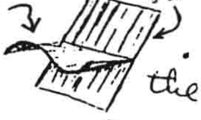


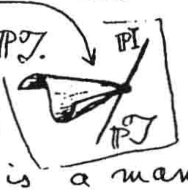
i.e. by

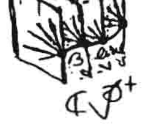
$$(\omega^A, \pi_{A'}) \rightsquigarrow (\omega^A \pi^{B'}, \pi_{A'}\pi^{B'}) = (\omega^{AB'}, \pi_{A'}{}^{B'})$$

$$\text{where } \omega^{AA'}\pi_{A'}{}^{B'} = 0,\ \pi_{A'}{}^{A'} = 0.$$

The space $\mathbb{T}^\#$ of blown up twistors is the space of such $(\omega^{AB'}, \pi_{A'}{}^{B'})$. For the space $\mathbb{PT}^\#$ of projective blown up twistors, we factor out by an overall factor. $\mathbb{PT}^\#$ is a (non-singular) complex 3-manifold in $\mathbb{CP}^6$. The projective blown-up twistors (elements of $\mathbb{PT}^\#$) represent α-lines on $\mathbb{C}\mathscr{I}^+$ together with their limits, the generators of $\mathbb{C}\mathscr{I}^+$. This limiting procedure is achieved by applying $\gamma\to 0$ to $(\gamma^{-1}\omega^A, \gamma\pi_{A'})$, which leaves us with blown-up twistors of the form $(\omega^A\pi^{A'}, 0)$, where $\pi^{A'}$ specifies a β-plane and ω^A an α-plane on $\mathbb{C}\mathscr{I}^+$, these intersecting in the generator in question. The α-lines and generators are scaled in terms of parallelly propagated covectors "pointing along" them (a conformally invariant notion), which gives the scalings needed to define non-projective blown-up twistors (elements of $\mathbb{T}^\#$).

In the curved case, this definition in terms of scaled α-lines and scaled generators works just as well as in the flat case, and provides us with the asymptotic (blown-up) twistor space $\mathcal{T}^\#$ (provisional definition). Dropping the scaling, we get $\mathbb{P}\mathcal{T}^\#$. It is not hard to see that $\mathbb{P}\mathcal{T}^\#$ is a non-singular complex manifold. The trouble, however, is that now the (googly) information of the SD radiation field seems to have got lost. This information lies in the location of the β-lines on $\mathbb{C}\mathscr{I}^+$. A β-line can be represented by the family of α-lines which meet it, i.e. by the corresponding "crinkly cone" in $\mathbb{P}\mathcal{T}$. The fact that these crinkly cones cannot be consistently represented as having flat tangent planes at their vertices on $\mathbb{PI}$ is a manifestation of the fact that $\mathbb{PI}$ is singular for $\mathbb{P}\mathcal{T}$. When we blow-up, we get smooth surfaces looking like  What is being proposed here is that if we specify the appropriate structure on $\mathbb{P}\mathcal{T}$ or $\mathbb{P}\mathcal{T}^\#$, then these surfaces will be singled out. It is not expected that this structure will manifest itself on $\mathcal{T}^\flat$, but rather in the way that $\mathbb{I}$ (or whatever is appropriate) is attached to $\mathcal{T}^\flat$ to yield $\mathcal{T}^\#$.





In $\mathbb{TN}$ 23(1-4), the forms

$$\delta = \overline{Z\,dZ} \qquad \text{and} \qquad \mathcal{D} = \frac{1}{6}\,\overline{Z\,dZ\wedge dZ\wedge dZ}\,,$$

both well defined on $\mathcal{T}^b$, with

$$\delta\wedge\mathcal{D}=0\,,\quad \delta\wedge d\delta=0\,,$$

were considered. It was noted that in the flat case the quantity

$$\Gamma=\delta\otimes\mathcal{D}$$

(a ⊞-tableau quantity) is regular (and non-zero) on $\mathbb{T}^\#$, whereas $\mathcal{D}$ blows up at $\mathbb{I}$ and δ goes to zero there. We can see this by choosing local coordinates ($w=0$ giving $\mathbb{I}$):

$$w=(Z^2)^2=\pi_{0'}\pi_{0'}\,,\quad x=\frac{Z^1}{Z^0}=\omega^1/\omega^0\,,\quad y=Z^0Z^2=\omega^0\pi_{0'}\,,\quad z=\frac{Z^3}{Z^2}=\pi_{1'}/\pi_{0'}$$

whence

$$\delta=w\,dz\,,\quad \mathcal{D}=\frac{y}{w}(w\,dy-y\,dw)\wedge dx\wedge dz\,,$$

$$\Gamma=dz\otimes dz\wedge(w\,dy-y\,dw)\wedge y\,dx\wedge dz\,.$$

In $\mathbb{T}$N 23, the hope was expressed that some differential operator might be found, which is coordinate independent for objects of the nature of Γ, and whose kernel would, in some appropriate sense, be objects defined by twistor functions homogeneous of degree -6. In fact, there _is_ such an operator, defined as follows. I shall use the notation $\circledwedge$ for a product of a 2-form with an n-form defined by

$$(dp\wedge dq)\circledwedge\alpha=dp\otimes dq\wedge\alpha-dq\otimes dp\wedge\alpha$$

and extended by linearity to arbitrary 2-forms. Then the required operator is given by

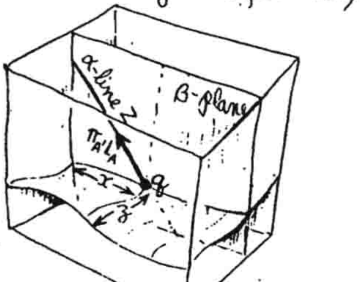
$$D(\delta\otimes\mathcal{D})=d\delta\circledwedge\mathcal{D}-\delta\otimes d\mathcal{D}$$

whence (by $\delta\wedge\mathcal{D}=0$)

$$D((\lambda\delta)\otimes\mathcal{D})=D(\delta\otimes(\lambda\mathcal{D})).$$

It follows that D does indeed operate on Γ itself, not depending on the particular way in which it is represented. $D\Gamma$ is a ⊞-tableau tensor. We also have $D(f\Gamma)=0$ iff f has homogeneity degree -6. We might hope to see a role for this in the encoding of googly information into the structure of $\mathcal{T}^\#$, but this has so far remained somewhat elusive. (For example, we might have found that there is some ambiguity in Γ in its natural extension to $\mathbb{I}$, but that $D\Gamma$ is defined unambiguously there.) However, Γ appears also to be unambiguously regular on $\mathcal{T}^{\#}$ (as here defined) in the curved case.

To see this, let us go to $\mathbb{C}\mathscr{I}^+$, a suitable conformal factor Ω having been chosen to make the metric and ε-spinors finite. We can (locally) make the induced metric on $\mathbb{C}\mathscr{I}^+$ flat and (locally) choose spinors ι^A, $\tilde{\iota}^{A'}$ constant on $\mathbb{C}\mathscr{I}^+$; where the vectors $\iota^A\tilde{\iota}^{A'}$ are tangent to the generators of $\mathbb{C}\mathscr{I}^+$. However, the spinors o^A, $\tilde{o}^{A'}$, needed to complete a spinor basis at each point of $\mathbb{C}\mathscr{I}^+$, cannot be chosen constant unless the Bondi-Sachs news function vanishes. There is curvature in the connection along α-planes on $\mathbb{C}\mathscr{I}^+$ if there is SD gravitational radiation and along β-planes if there is ASD gravitational radiation. Choose (locally) an arbitrary cut $\mathscr{C}$ of $\mathbb{C}\mathscr{I}^+$ (smooth cross-section of the generators) and mark the point q where an α-line or generator intersects $\mathscr{C}$. We adopt a local twistor description of the α-lines, so that along each α-line a twistor corresponding to that α-line has the description



$$Z^\alpha = (0, \pi_{A'}),$$

each tangent vector the α-line being proportional to $\iota^A\pi^{A'}$. (It is important to bear in mind that, because of the conformal rescaling, this "$\pi_{A'}$" is quite different from the "$\pi_{A'}$" used as a flat-twistor-space coordinate above, where we had (old $\pi_{A'}$) $\overline{Z} = (0, \pi^{A'})$.) We recall (R.P. & W.R. Spinors & Space-Time Vol. 2, pp. 376, 65; $\mathbb{T}\mathbb{N}$ 23, p2) that on $\mathbb{C}\mathscr{I}^+$ the infinity twistor takes the form

$$\Pi = I_{\alpha\beta} = \begin{bmatrix} 0 & -i\,\iota_A\tilde{\iota}^{A'} \\ i\,\tilde{\iota}^{A'}\iota_B & 0 \end{bmatrix}$$

($\rho' = 0$, $A = 1$, being chosen here, in the notation of S.&S-T.). Thus $Z\rceil$ has the local twistor description

$$Z^{\alpha} I_{\alpha\beta} = \left(\iota_B (i \tilde{\iota}^{A'} \pi_{A'}),\ 0 \right) = i\pi_{1'} (\iota_B,\ 0),$$

so the covector associated with $\mathrm{I}Z\rceil$ is

$$\mathscr{P}_a = (i\,\pi_{1'})\, \iota_A \pi_{A'} .$$

Local coordinates for $\mathscr{T}^{\#}$ (for the α-line non-tangent to $\mathscr{C}$) can be taken to be the two independent components of $\mathscr{P}_a$ and the two coordinates (x, y) specifying the point q on $\mathscr{C}$. (See figure on previous page). Indeed, we obtain local coordinates (w, x, y, z), closely analogous to (and direct generalizations of) those introduced for $\mathbb{T}^{\#}$ above, by putting

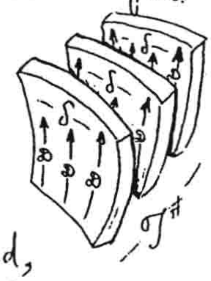
$$w = i\,\pi_{1'}\pi_{1'}\ , \quad y = i\,\pi_{1'}\pi_{0'}\ ,$$

the ratio $w/y = \pi_{1'}/\pi_{0'}$ of these two quantities being a coordinate defining the direction of the x-line through q (in the β-plane on $\mathbb{C}\mathscr{I}^{+}$ through q). Provided that the scalings for z and x are chosen appropriately in relation to ι^A and $\tilde{\iota}^{A'}$, we find that, just as before,

$$\delta = y\,dz \quad \text{and} \quad \mathscr{D} = \frac{y}{w}(w\,dy - y\,dw) \wedge dx \wedge dz .$$

It follows that there can be no "googly information" solely in the behaviour of the forms δ and $\mathscr{D}$ on $\mathscr{T}^{\#}$, with the definition of $\mathscr{T}^{\#}$ so far given. (It is of interest to note that the ratio y/w of the $\pi_{A'}$ components used here define the slope of the α-line, with $\pi_{1'} = 0$ for a generator. This may be contrasted with the "standard" $\pi_{A'}$ used earlier, for flat twistor space $\mathbb{T}$, the ratio of whose components fixed the choice of β-plane, this ratio being essentially the quantity z.

The geometrical meaning of the tensor $\mathbf{\Gamma}$ up to proportionality — equivalently, the pair of forms δ and $\mathscr{D}$ up to proportionality — is that it defines a structure on $\mathscr{T}^{\#}$ that is the same at each point. This structure is a foliation by curves (the integral curves of the Euler vector field Υ, defined as the dual of $\mathcal{D}$, dualized with respect to $\frac{1}{4}d\mathcal{D}$) which lie in 3-surfaces (whose tangent spaces are duals to δ). The Frobenius relation $\delta_\wedge d\delta = 0$ ensures that these 3-surfaces are locally integrable. Moreover $\delta_\wedge \mathcal{D} = 0$, as we have here, is the condition that the curves lie on the 3-surfaces.



As a point of twistor philosophy, one might take the view that all information should be stored in global structure, the neighbourhood of any point being indistinguishable from that of the neighbourhood of any other. It is not clear how closely such a viewpoint can be followed, but the use of the space $\mathcal{T}^\#$ is in accordance with its spirit. We note that even in the flat case the non-blown-up space $\mathbb{T}$ does not qualify, since the points of I do not have neighbourhoods with the above structure, whereas the blown-up space $\mathbb{T}^\#$ does.

In the standard nonlinear graviton construction, one uses slightly more than the above local structure since the space-time metric itself, as opposed to merely the conformal metric, uses the forms δ and $\mathcal{D}$ themselves and not just their proportionality classes. More precisely, it is the ratio $\mathcal{D}/\delta\otimes\delta$ that comes in to define the metric (for δ defines $I_{\alpha\beta}$ which is, in effect, $\varepsilon^{A'B'}$ and $\delta/\mathcal{D}$ defines $I^{\alpha\beta}$ which is, in effect, ε^{AB}; thus $\delta\otimes\delta/\mathcal{D}$ in effect defines g^{ab}). We may imagine assigning a new $\hat{\delta}$ and $\hat{\mathcal{D}}$ in the neighbourhood of some point of $\mathcal{T}^\flat$, compatible with the foliation structure, so $\hat{\delta}$ and $\hat{\mathcal{D}}$ are scalar multiples of δ and $\mathcal{D}$. For the metric of $\mathcal{M}$ to be preserved, we must have

$$\hat{\delta} = \chi\delta, \quad \hat{\mathcal{D}} = \chi^2\mathcal{D},$$

whence

$$\hat{\Gamma} = \chi^3\Gamma.$$

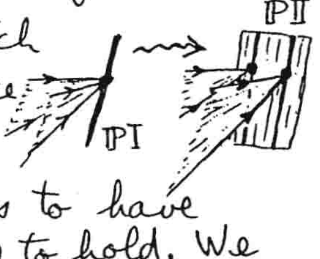
If we allow for a conformal factor (needed for $\mathbb{C}\mathscr{I}^+$, in any case), so that

$$\hat{g}_{ab} = \Omega^2 g_{ab},$$

we have
$$\hat{\delta}=\Omega^{-1}\chi\,\delta,\quad \hat{\mathfrak{D}}=\chi^{2}\mathfrak{D},\quad \hat{\mathbb{\Gamma}}=\Omega^{-1}\chi^{3}\,\mathbb{\Gamma}$$
corresponding to $\hat{\varepsilon}_{AB}=\Omega\chi\,\varepsilon_{AB}$, $\hat{\varepsilon}_{A'B'}=\Omega\chi^{-1}\varepsilon_{A'B'}$. However, we should bear in mind that these rescalings apply at points of $\mathcal{T}^{\#}$ rather than at individual points of $\mathcal{M}$. (A conformal rescaling of $\mathcal{M}$ does not preserve δ, even up to proportionality; the rescaled $\mathcal{M}$ would not be Ricci-flat.)

Now suppose that a rescaling is introduced which makes both of $\hat{\delta}$ and $\hat{\mathfrak{D}}$ regular and non-zero at $\mathbb{I}$. As things stand, there is too much choice for such scalings. We want somehow to restrict this choice so that the googly information is encoded. It may be noted that some choices are distinguished by the fact that they yield a 2-form δ that is closed, so locally
$$\hat{\delta}=d\zeta.$$
In the flat case, we may take
$$\Omega\chi^{-1}=\omega,\quad \zeta=z$$
and we recall that $z=\pi_{1'}/\pi_{0'}$, where $\pi_{0'},\pi_{1'}$ are the ordinary flat-space $\pi_{A'}$-components. $SL(2,\mathbb{C})$ transformations of these coordinates are generated by the two types of replacement
$$\zeta\longmapsto\zeta+\text{const.},\quad \zeta\longmapsto-\zeta^{-1}.$$
For the first of these, the rescaling factor $\Omega^{-1}\chi$ is unchanged; for the second, it is replaced by $\Omega^{-1}\chi\zeta^{-2}$. This procedure applies to any rescaling factor for which $\hat{\delta}$ is closed, so each such choice provides us with a local (primed) spin-space. As described in $\mathbb{T}$N 23, this family of primed spin-spaces encodes, in a sense, the entire googly information. Thus, we

require to know the appropriate "preferred" rescaling factors. Each such factor would have to depend upon the actual direction in which a point of $\mathbb{P}I$ is approached, from within $\mathbb{P}\mathcal{T}^b$, which still leaves some directional dependence at individual points of $\mathbb{P}II$: $\rightsquigarrow$



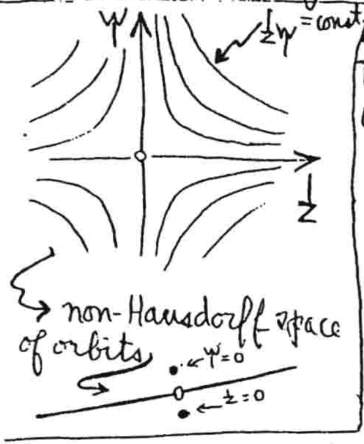
The rescaling factor $\Omega^{-1}X$ has to have homogeneity -2 in order for $d\hat{\delta}=0$ to hold. We are not at liberty to incorporate twistor scale-dependence into the conformal factor Ω, so taking that dependence to lie in X, we find that the homogeneity degree for the scale factor for $\hat{\Gamma} = \Omega^{-1}X^3\,\Gamma$ is -6. This has some suggestiveness for a role for the googly twistor function, but the matter is not entirely clear, as yet.

In any case, we still have the problem of fixing the appropriate scale factors in terms of the complex geometry of $\mathcal{T}^\#$. As specified so far, the geometry of $\mathcal{T}^\#$ does not seem to contain this information, and a further idea is needed. From the twistor point of view, there is something a little unnatural about forming the space of products $\overset{|}{Z}\bar{Z}\rceil$ to replace the space of twistors $\overset{|}{Z}$. I wish to adopt a subtly different viewpoint here. We are to think of the product $\overset{|}{Z}\bar{Z}\rceil$ as a particular case of a product $\overset{|}{Z}\underset{|}{W}$, where the dual twistor $\underset{|}{W}$ specifies a covector $\underset{dZ}{\overset{W}{|}}$. The covector bundle $T^*\mathcal{T}^b$ of $\mathcal{T}^b$ is, of course, well defined; and one can easily specify a covector at a point $\overset{|}{Z}$ of $\mathcal{T}^b$ in terms of $\mathbb{C}\mathscr{I}^+$, where $\underset{|}{W}$ is represented as a local dual twistor which is local-twistor constant along the α-line representing $\mathbb{P}\overset{|}{Z}$. The exterior derivative $\underset{\smile}{dW\wedge dZ})$ of $\underset{dZ}{\overset{W}{|}}$ is the

2-form defining the symplectic structure of $T^*\mathcal{T}^b$. We are interested in the subspace of $T^*\mathcal{T}^b$ defined by $\frac{\psi}{z}=0$ (or possibly $\frac{\psi}{z}=k$ – of relevance to APH's work?). Taking $\frac{\psi}{z}$ as a Hamiltonian, we obtain the flow whose orbits are to be factored out by in order to obtain the relevant "reduced phase space", where

$$(\lambda\psi, \tfrac{1}{z}) \equiv (\psi, \lambda\tfrac{1}{z})$$

(integral curves of $\frac{\psi}{\partial_\psi} - \frac{\partial_z}{z}$). The product $\frac{1}{z}\psi$ can be used to represent these equivalence classes, for the most part, but there is a subtle difference. There are those particular orbits for which either $\frac{1}{z}=0$ or $\psi=0$, in addition, giving a reduced phase space which is *non-Hausdorff*. In fact, for an ordinary cotangent bundle, one may not take $\frac{1}{z}=0$, whereas the space given by $\psi=0$ is the canonical copy of the original (twistor) space as a subspace of its cotangent bundle. Indeed, it is this zero-space that enables us to reconstruct the 1-form $\frac{\psi}{dz}$ from the 2-form $d\psi\wedge dz$) in a unique way. To give meaning to the subspace $\frac{1}{z}=0$, it would be necessary to have some identification between covectors at *different* $\frac{1}{z}$-points as being, in an appropriate sense, the *same* ψ.


ψ
$\frac{1}{z}\psi=\text{const.}$
$\frac{1}{z}$
non-Hausdorff space of orbits
$\psi=0$
$\frac{1}{z}=0$

One "solution" to the googly problem, therefore, would be to attach a "$\frac{1}{z}=0$ zero-set" to the cotangent bundle $T^*\mathcal{T}^b$. It seems to me, however, that this procedure would be more in the nature of "begging the question" than in providing a solution to the googly problem. It would really be stretching a point to consider this procedure to be providing us with an extended "twistor space". The resulting space would be completely symmetrical with respect to twistors and dual twistors and is really much more related to ambitwistor space than to either of $\mathcal{T}$ or $\mathcal{T}^*$ individually.

Nevertheless, I believe that the above ideas provide an appropriate setting for what one should really be doing to construct $\mathcal{T}^\#$. What is required, initially, is the sub-bundle of $T^*\mathcal{T}^\flat$ for which the covector ψ has the special form $x\rceil$. The "Hamiltonian" $\overset{W}{\underset{z}{|}}$ now becomes $\overline{x\,z}$; accordingly we appear to have to factor out by

$$(\lambda\, x\rceil,\ \overset{|}{z}) \equiv (x\rceil,\ \lambda\,\overset{|}{z})$$

where $\overline{x\,z}=0$, i.e. where $x\rceil$ and $z\rceil$ are proportional. To a first approximation, this indeed gives the blown-up twistor space $\mathcal{T}^\#$, as defined above. Ignoring the issue of zero-sets, for the moment, we can think of a point of this space as having elements of the form

$$\overset{|}{z}\,x\rceil,$$

with $x\rceil$ proportional to $z\rceil$, where $\overline{x\,dz}$ is a covector at the point $\overset{|}{z}$. We have the freedom to choose the proportionality factor by putting it equal to unity, so our space (the $\mathcal{T}^\#$ defined above) consists of objects

$$\overset{|}{z}\,z\rceil,$$

where $z\rceil$ defines a (restricted) covector:

$$\delta = \overline{z\,dz}\,.$$

However, we need to consider the zero sets, and it seems to me that the googly information ought to be contained therein. Rather than merely considering the two alternatives $X\rceil = 0$ and $\overset{|}{Z} = 0$ in $(\lambda X\rceil, \overset{|}{Z}) \equiv (X\rceil, \lambda \overset{|}{Z})$ with $\overline{X Z} = 0$, it may be that we ought to view all this slightly differently. The space of $(X\rceil, \overset{|}{Z})$ is not, after all, a non-degenerate symplectic manifold, the 2-form $d\overline{X_\wedge dZ}$ having rank 4 on a 6-dimensional space. For a start, we may examine the zero-sets in relation to the "π-spaces" $(X\rceil, Z\rceil)$ alone. Putting (flat case) $\Psi = X\rceil = (0, \xi^{A'})$, $\overset{|}{Z} = (\omega^A, \pi_{A'})$, we have

$$d\Psi_\wedge dZ = d\overline{X_\wedge dZ} = d\xi^{A'}{}_\wedge d\pi_{A'}.$$

Taking the Hamiltonian to be $\xi^{A'}\pi_{A'}$ $(= \overline{X Z})$, we are led to contemplate the equivalence

$$(\lambda\xi^{A'}, \pi_{A'}) \equiv (\xi^{A'}, \lambda\pi_{A'})$$

with zero-sets given by $\xi^{A'} = 0$ and $\pi_{A'} = 0$. However, all of this is parameterized by ω^A in some way. It should be noted that the limits like

$$(\omega^A, \lambda\pi_{A'}) \to (\omega^A, 0)$$

given when $\lambda \to 0$ are just those of the kind in which an α-line rotates about a point of $\mathbb{C}\mathscr{I}^+$ to become a generator, as considered earlier. More work needs to be done to discover the detailed nature of the appropriate zero-sets and how quantities like Γ are to be defined and extended to these zero-sets. The condition that such extensions to the zero-set regions exist could well provide the necessary restrictions on, say, Γ that encode the full googly information — and perhaps, if this is successful, the required source information also. Work very much in progress.....

Thanks to L.M.J., E.T.N., J.F., ...

250

INCIDENCE BETWEEN COMPLEX NULL RAYS

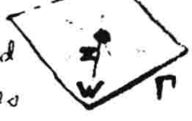
Twistor Newsletter **41** (1996)

This topic is developed also in Chapters 255 and 260.

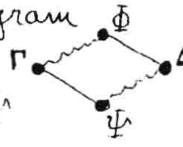
Incidence between Complex Null Rays

The incidence relations between complex null rays exhibit certain curious properties. I shall try to illuminate these here. The structures that arise have relevance to the geometry of ambitwistor space. The discussion will be phrased in terms of compactified complex Minkowski space $\mathbb{C}\mathbb{M}^{\#}$ and its associated (projective) twistor space $\mathbb{P}\mathbb{T}$, in the first instance, with some generalizations to curved space-time indicated at the end.

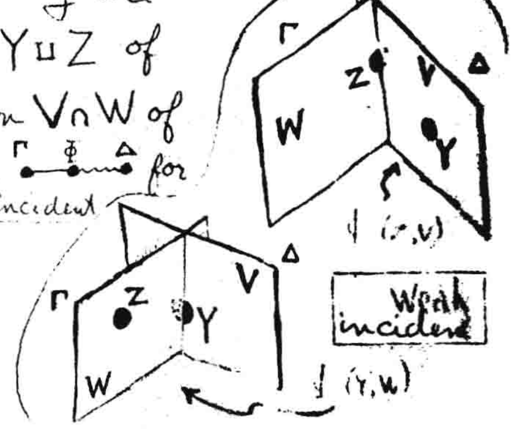
By a *ray* I mean a (generally complex) null geodesic. A ray γ in $\mathbb{C}\mathbb{M}^{\#}$ has a twistor interpretation in terms of a pair $\Gamma = (Z, W)$, where $Z \in \mathbb{P}\mathbb{T}$, $W \in \mathbb{P}\mathbb{T}^{*}$ with Z incident with W (i.e. $Z^{\alpha} W_{\alpha} = 0$). I shall refer to Γ as a *pencil*, in reference to the plane pencil of lines lying in the plane $W \subset \mathbb{P}\mathbb{T}$ and passing through the point $Z \in \mathbb{P}\mathbb{T}$, these lines representing the points of the ray $\gamma \subset \mathbb{C}\mathbb{M}^{\#}$. The point $Z \in \mathbb{P}\mathbb{T}$ is called the *vertex* of the pencil and W is its plane.

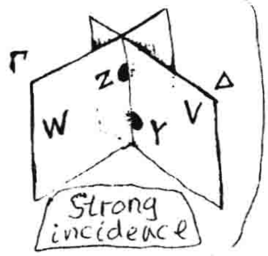


There are various kinds of incidence relation holding between rays in $\mathbb{C}\mathbb{M}^{\#}$, which I shall call "weak incidence", "strong incidence", "α-incidence", and "β-incidence". Two pencils $\Gamma = (Z, W)$ and $\Delta = (Y, V)$ are *α-incident* if $Y = Z$ (diagram $\overset{\Gamma}{\bullet}\!\sim\!\sim\!\sim\!\overset{\Delta}{\bullet}$); they are *β-incident* if $V = W$ (diagram $\overset{\Gamma}{\bullet}\!-\!-\!-\!\overset{\Delta}{\bullet}$); they are *weakly incident* if the join $Y \sqcup Z$ of Y and Z meets the intersection $V \cap W$ of V and W (diagram $\overset{\Gamma}{\bullet}\!\sim\!\sim\!\overset{\Phi}{\bullet}\!-\!-\!\overset{\Delta}{\bullet}$ or $\overset{\Gamma}{\bullet}\!-\!-\!\overset{\Phi}{\bullet}\!\sim\!\sim\!\overset{\Delta}{\bullet}$ for some Φ); they are *strongly incident* if $Y \sqcup Z = V \cap W$ (diagram



for some Φ, Ψ).

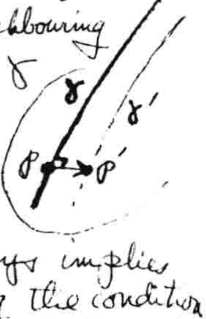




Strong incidence between Γ and Δ is the condition that the corresponding rays γ and δ in $\mathbb{C}\mathbb{M}^{\#}$ have a point in common. (See also the discussion by TRF in TN 38 pp. 30–34.)

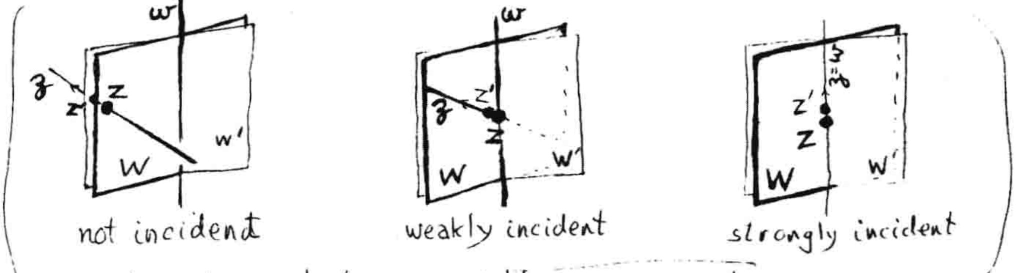
Note that the space $\mathfrak{D}$ of those Δ that are weakly incident with a fixed Γ is the union of two smooth manifolds $\mathfrak{D}_\alpha$, $\mathfrak{D}_\beta$ (respectively), each being a $\mathbb{CP}^2$ bundle over $\mathbb{CP}^2$ This arises because each of the spaces of pencils Φ α-incident or β-incident with Γ is a $\mathbb{CP}^2$, and we apply one step in each to arrive at weak incidence. The condition of strong incidence does not provide a smooth manifold, however, as we shall be seeing shortly.

Consider, next, weak and strong incidence with γ of a ray γ′ neighbouring to γ (i.e. we are concerned with infinitesimal variations of rays away from γ). We find that the notion of weak incidence between γ′ and γ is equivalent to abreastness. We say that γ′ is abreast with γ if a connecting vector from a point p on γ to a neighbouring point p′ of γ′ is orthogonal to the direction of γ (cf. Penrose & Rindler, Spinors & Space-time Vol. 2 p. 176). This property is independent of the choice of p on γ and of the neighbouring point p′ on γ′. An easy way to see that weak incidence between neighbouring rays implies abreasness is to observe that a way of stating the condition for two rays γ and δ to be weakly incident is that there be a ray meeting each of γ and δ orthogonally. This is the ray φ represented by the pencil Φ in $\overset{\Gamma}{\bullet}\!\sim\!\overset{\Phi}{\bullet}\!-\!\overset{\Delta}{\bullet}$ or $\overset{\Gamma}{\bullet}\!-\!\overset{\Phi}{\bullet}\!\sim\!\overset{\Delta}{\bullet}$. (Note that any two rays in the same α-plane or in the same β-plane are orthogonal at their intersection point.) Hence, when δ becomes γ′, neighbouring to γ, with p the intersection of φ with γ, we see that the condition for abreasness is satisfied. For the converse, we may simply note that weak incidence is a codimension 1 condition, so the neighbourhood of $\mathfrak{D}$ at Γ cannot be smaller than the space of γ′ abreast



with γ," abreastness being an irreducible condition.

In terms of $\mathbb{PT}$, weak incidence between Γ and Γ' is understood as the condition that the join $z = Z \sqcup Z'$ meet the intersection $w = W \cap W'$. Unlike the case **of finitely** differing pencils, the condition now becomes symmetrical, w and z just being two lines in the plane **W** passing through the point **Z**.

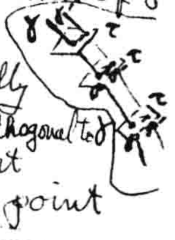


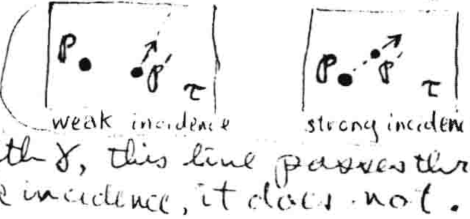
not incident weakly incident strongly incident

[Incidence between neighbouring pencils.

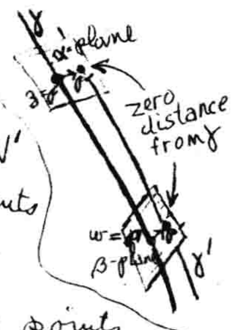
What about strong incidence? This is the condition that γ' actually "meets" γ', so that the pencils Γ and Γ' have a line in common. This line must be both z and w, so we have $z = w$ as the condition for strong incidence. Recall that with finitely differing pencils Γ, Δ, **if both** forms of weak incidence hold together, then strong incidence must hold. It is curious that for infinitesimally differing pencils Γ, Γ', the situation is quite different. The two forms of weak incidence for Γ, Γ' coincide, and the simultaneous holding of both forms is insufficient to imply strong incidence. I shall come to the explanation of this curious fact shortly.

In terms of space-time geometry, we can think of γ' as given by a Jacobi field of connecting vectors along γ. Referring this to a family of parallelly propagated 2-plane elements τ, at the points of γ, orthogonal to γ, all identified with one another by parallel transport along γ, we find that, for an abreast ray γ', the point ρ' in this plane executes a straight line in τ, uniformly with the affine parameter on γ. If γ' is strongly incident with γ, this line passes through **ρ**; in the general case of weak incidence, it does not.





weak incidence strong incidence



In the case of strong incidence, the point where ρ' coincides with ρ is the point $z \in \gamma$, represented by the line $z = Z \sqcup Z' = w = W \cap W'$ in $\mathbb{PT}$. In the general case of weak incidence, the points z and w are distinct, and they represent the locations of ρ at which the point ρ' reaches zero distance from ρ. These are imaginary points, when the rays are real, and they occur where γ' intersects the α-plane and the β-plane through γ, respectively.

In terms of twistor notation, we can write

$$Y^\alpha = Z^\alpha + \delta Z^\alpha, \quad V_\alpha = W_\alpha + \delta W_\alpha$$

where

$$Z^\alpha W_\alpha = 0, \quad Y^\alpha V_\alpha = 0$$

whence

$$W_\alpha \delta Z^\alpha + Z^\alpha \delta W_\alpha + \delta Z^\alpha \delta W_\alpha = 0.$$

To first order, we have

$$W_\alpha \delta Z^\alpha = -Z^\alpha \delta W_\alpha.$$

The condition for weak incidence is either $Z^\alpha V_\alpha = 0$, i.e.

$$Z^\alpha \delta W_\alpha = 0$$

or $Y^\alpha W_\alpha = 0$, i.e.

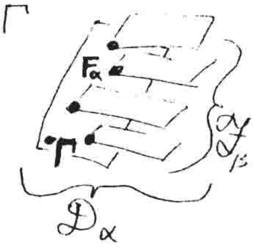
$$W_\alpha \delta Z^\alpha = 0$$

which, to first order, are indeed equivalent. To recognise strong incidence we must go to second order, the condition being that, in addition, the quadratic relation

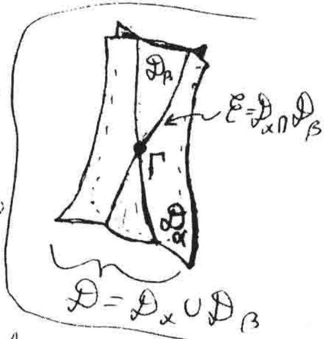
$$\delta Z^\alpha \delta W_\alpha = 0$$

must hold. (For real light rays the discussion is the same, but with $W_\alpha = \bar{Z}_\alpha$, $\delta W_\alpha = \delta \bar{Z}_\alpha$.)

Let us try to understand this curious situation in terms of the geometry of ambitwistor space $\mathbb{A}$, for $\mathbb{CM}^\#$. ($\mathbb{A}$ is the space of Γs, i.e. of complex rays in $\mathbb{CM}^\#$.) This space is **5**-dimensional. The space $\mathfrak{D}$ of points of $\mathbb{A}$ that are weakly incident with Γ is, as we have seen, the union of two smooth manifolds $\mathfrak{D}_\alpha$ and $\mathfrak{D}_\beta$. We can think of $\mathfrak{D}_\alpha$ as constructed as follows. There is a foliation $\mathcal{F}_\alpha$ of $\mathbb{A}$ by families of points of $\mathbb{A}$ that are α-incident with each other and another foliation $\mathcal{F}_\beta$ by families of points of A that are β-incident with each other. To construct $\mathcal{D}_\alpha$, we fix the member F_α of $\mathcal{F}_\alpha$ through Γ and sweep out a region of A by taking all the members of $\mathcal{F}_\beta$ through points of F_α. In an exactly similar way, we can construct $\mathcal{D}_\beta$, now by fixing F_β through Γ and allowing the $\mathcal{F}_\alpha$ members through points of F_β to vary. Note that the difference between $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$, when constructed in this way is a commutator, and this difference is thus of second order. In other words, $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$ have the same tangent plane at Γ. This explains the symmetry between the two forms of weak incidence when we look only to first order.



However, the strong incidence locus $\mathcal{E}$ is the intersection of $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$. Since $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$ touch at Γ, their intersection has a conical singularity at Γ. The equation, locally, of this cone is the quadratic relation $\delta Z^\alpha \delta W_\alpha = 0$.


$\mathcal{D}_\beta$
$\mathcal{E} = \mathcal{D}_\alpha \cap \mathcal{D}_\beta$
Γ
$\mathcal{D}_\alpha$
$\mathcal{D} = \mathcal{D}_\alpha \cup \mathcal{D}_\beta$

These ideas can also be applied in a general curved space-time $\mathcal{M}$ in reference to a choice of hypersurface $\mathcal{H}$ (possibly $\mathcal{H} = \mathscr{I}^+$, for example). In this case $\mathbb{P}\mathbb{T}$ and $\mathbb{P}\mathbb{T}^*$ are, respectively, the α-curves and β-curves in $\mathbb{C}\mathcal{H}$. Incidence between α-curves and β-curves is the condition that they intersect. The other notions of incidence are built up from this. The geometry described above is virtually unchanged. (Of course all this is relative to the choice of $\mathcal{H}$. If $\mathcal{H}$ is moved within $\mathcal{M}$, the structure changes.)

Further details are to appear in an article in honour of Andrzej Trautman's 64th birthday.

251
TWO PUZZLES

Twistor Newsletter **41** (1996)

This also contains the solution to the puzzle in Chapter 118.

Two Puzzles

① In T N 10, p.22 (July 1980), I proposed, as a puzzle, to find (?) in the following sequence:

…, 7, 9, 12, (?), 24, 36, 56, 90, …

The point about the answer

$$\text{(?)} = 24 \log 2 = 16{\cdot}6355323\ldots$$

is that it is transcendental, being given (by use of l'Hospital's rule) by $n=0$ in

$$\frac{24}{n}(2^n - 1)$$

all other terms in the sequence being rational numbers.

Now try

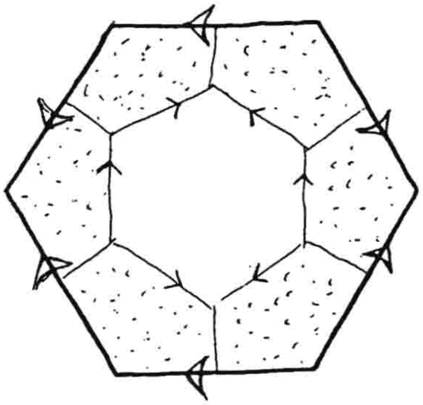
…, 28, 0, 21, 4, 18, 0, (?), 24, 18, 20, 21, 24, 28, …

which has a little extra twist about it.

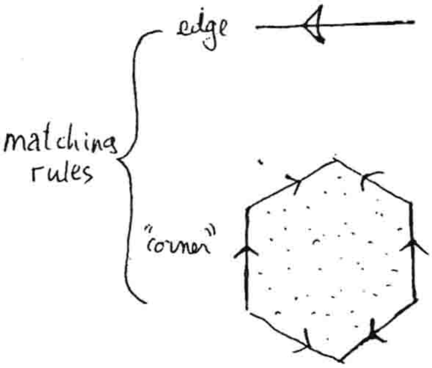
[Hint: first add a certain integer to each term, to make it somewhat more symmetrical]

---

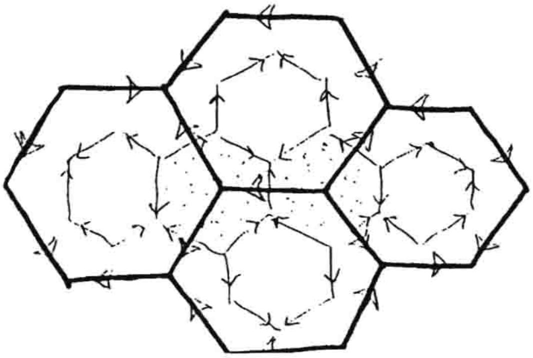
② This problem is to tile the plane with a regular hexagon, with edge-matching rules and "corner matching rules":



tile



To start:



252

ON GRAVITY'S ROLE IN QUANTUM STATE REDUCTION

Gen. Relativity Gravitation **28** (1996), no. 5, 581–600

This paper provides a physical justification of the 'Diósi–Penrose proposal' for quantum state reduction (see Chapters 215 and 218), based on a *tension* between Einstein's principle of general covariance and the quantum-mechanical principle of linear superposition. See also Chapter 253.

On Gravity's Role in Quantum State Reduction

Roger Penrose[1,2]

Received August 22, 1995. Rev. version December 12, 1995

The stability of a quantum superposition of two different stationary mass distributions is examined, where the perturbing effect of each distribution on the space-time structure is taken into account, in accordance with the principles of general relativity. It is argued that the definition of the time-translation operator for the superposed space-times involves an inherent ill-definedness, leading to an essential uncertainty in the energy of the superposed state which, in the Newtonian limit, is proportional to the gravitational self-energy E_Δ of the difference between the two mass distributions. This is consistent with a suggested finite lifetime of the order of $\hbar/E_\Delta$ for the superposed state, in agreement with a certain proposal made by the author for a gravitationally induced spontaneous quantum state reduction, and with closely related earlier suggestions by Diósi and by Ghirardi *et al.*

1. THE PROBLEM OF QUANTUM STATE REDUCTION

The fundamental problem of quantum mechanics, as that theory is presently understood, is to make sense of the *reduction of the state vector* (i.e. *collapse of the wavefunction*), denoted here by R. This issue is usually addressed in terms of the "quantum measurement problem", which is to comprehend how, upon *measurement* of a quantum system, this (seemingly) discontinuous R-process can come about. A measurement, after all, merely consists of the quantum state under consideration becoming entangled with a more extended part of the physical universe, e.g. with a

---

[1] Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK

[2] Center for Gravitational Physics and Geometry, Pennsylvania State University, University Park, Pennsylvania 16801, USA

measuring apparatus. This measuring apparatus — together with the observing physicist and their common environment — should, according to conventional understanding, all also have some quantum description. Accordingly, there should be a quantum description of this *entire* quantum state, involving not only the original system under consideration but also the apparatus, physicist, and remaining environment — and this entire state would be expected to evolve continuously, solely according to the Schrödinger equation (unitary evolution), here denoted by the symbol U.

Numerous different attitudes to R have been expressed over many years, ever since quantum mechanics was first clearly formulated. The most influential viewpoint has been the "Copenhagen interpretation" of Niels Bohr, according to which the state vector $|\psi\rangle$ is not to be taken seriously as describing a quantum-level physical reality, but is to be regarded as merely referring to our (maximal) "knowledge" of a physical system, and whose ultimate role is simply to provide us with a means to calculate probabilities when a measurement is performed on the system. That "measurement" would be taken to come about when that system interacts with a *classical* measuring apparatus. Since "our knowledge" of a physical system can undergo discontinuous jumps, there is no reason to be surprised — so it would be argued — when $|\psi\rangle$ undergoes discontinuous jumps also!

Closely related is the environment-entanglement "FAPP" ("for all practical purposes") point of view (e.g. Ref. 43; cf. Ref. 3), according to which the R-process is taken to be some kind of approximation to the U-evolution of the system together with its environment, and R is viewed as having taken place in the system itself "for all practical purposes" [3]. The essential idea is that the environment involves enormously numerous random degrees of freedom, and these become entangled with the limited number of degrees of freedom in the system itself. Accordingly, the delicate phase relations between the system's degrees of freedom become irretrievably lost in these entanglements with the environment. It is argued that this effective loss of phase coherence in the system itself gives rise to R, FAPP, although certain additional assumptions are needed in order for this conclusion to be entertained.

There are also viewpoints — referred to as the "many worlds", or "many minds", or the Everett interpretation (Ref. 13; cf. Ref. 10) — whereby it is accepted that all the different macroscopic alternatives which (according to U) must remain superposed, actually co-exist *in reality*. It can be convincingly argued that if $|\psi\rangle$ is indeed taken to represent an *actual* reality at all levels, and if $|\psi\rangle$ evolves *precisely* according to U, then something of this nature must hold. However, by itself, and without further assumptions, this would offer no explanation as to why the "illusion" of merely one world presents itself to our consciousnesses, nor why the correct quantum-mechanical probabilities come about (Refs. 38,39; cf. also Ref. 36, sec.6.2,6.7).

Other alternative interpretations have been put forward, such as those of de Broglie [9], Bohm [5], Bohm and Hiley [6], Haag [19], Griffiths [18], Omnès [25], and Gell-Mann and Hartle [15], in which the standard quantum procedures are reformulated in a different mathematical framework. The authors of these proposals do not normally take the view that any experimentally testable deviations from standard quantum mechanics can arise within these schemes.

Set against all these are proposals of a different nature, according to which it is argued that present-day quantum mechanics is a limiting case of some more unified scheme, whereby the U and R procedures are *both* to be approximations to some new theory of physical reality. Such a theory would have to provide, as an appropriate limit, something equivalent to a unitarily evolving state vector $|\psi\rangle$. Indeed, many of these schemes, such as that of Pearle [27,28], Bialynicki-Birula and Mycielski [4], Ghirardi, Rimini, and Weber [16], and Weinberg [41], use a quantum-state description $|\psi\rangle$, just as in standard quantum mechanics, but where the evolution of the state deviates by a tiny amount from the precise Schrödinger (or Heisenberg) evolution U. Related to these are proposals which posit that it is in the behaviour of *conscious beings* that deviations from precise U-evolution are to be found (cf. Ref. 42). But in the more "physical" such schemes, the suggested deviations from standard U-evolution become noticeable merely when the system becomes "large", in some appropriate sense. This "largeness" need not refer to physical dimension, but it might, for example, be the number of particles in the system that is relevant (such as in the specific scheme put forward in Ref. 16). In other proposals, it is considered that it is the mass, or mass distribution, that is all important. In schemes of the latter nature, it is normally taken that it is *gravity* that provides the influence that introduces deviations from the standard quantum rules. (See Refs. 20–24,11,17,29,37,30–36, for proposals of this nature.)

In this paper, a new argument is given that explicitly supports a gravitational role in state-vector reduction. Most particularly, it is consistent with a particular criterion, explicitly put forward in [35,36], according to which a macroscopic quantum superposition of two differing mass distributions is *unstable* (analogous to an unstable particle). Accordingly, such a state would decay, after a characteristic lifetime T, into one or the other of the two states. To compute T, we take the *difference* between the two mass distributions under consideration (so that one counts positively and the other negatively) and compute its gravitational self-energy E_Δ. The idea is that we then have

$$T \simeq \hbar / E_\Delta \, .$$

This criterion is also very close to one put forward earlier by Diósi [11], and as modified by Ghirardi, Grassi, and Rimini [17], but their point of view is different from the one used here, proposed by Penrose [34–36]. In particular, in their type of scheme, there are violations of energy conservation which could — in principle or in practice — be experimentally detected. With the type of proposal in mind here, the idea would be that such violations of energy conservation ought to be absent, owing to the fundamental involvement of some of the basic principles of general relativity. (There is a brief discussion of this point in Ref. 36, p.345.) In this article it is indicated that there is a basic conflict between Einstein's general covariance principle and the basic principles of quantum theory, as they relate to stationary states of superposed gravitational fields. It is argued that this conflict can be resolved within the framework of the specific state-reduction proposal of [35,36], according to which such superposed gravitational fields are essentially *unstable*.

It should be made clear, however, that this proposal does not provide a *theory* of quantum state reduction. It merely indicates the level at which deviations from standard linear Schrödinger (unitary) evolution are to be expected owing to gravitational effects. Indeed, it is this author's personal opinion that the correct theory uniting general relativity with quantum mechanics will involve a major change in our physical world-view — of a magnitude at least comparable with that involved in the shift from Newtonian to Einsteinian gravitational physics. The present paper makes no pretensions about even aiming us in the right direction in this regard. Its purpose is a different one, namely to show that even within the framework of completely conventional quantum theory, there is a fundamental issue to be faced, when gravitational effects begin to become important. Standard theory does not provide a clear answer; moreover, it allows room for the type of instability in superposed states that would be consistent with the proposals of Refs. 11,17 and 34–36.

2. STATIONARY STATES

Let us consider the following situation. We suppose that a quantum superposition of two states has been set up, where each individual state has a well-defined static mass distribution, but where the mass distributions differ from one state to the other. For example, we could have a rigid lump of material which we contrive to place in a quantum superposition of two different locations. Such a superposed state could be achieved by having a photon simultaneously transmitted through and reflected off a half-silvered mirror placed at a distance from the lump, where the transmitted part of the photon's state then triggers a device which slowly moves the lump from its initial location to somewhere nearby, but the reflected part leaves the lump alone. (This is an inanimate version of "Schrödinger's cat".)

If we ignore any gravitational effects, the two alternative locations of the lump will each be stationary states, and will therefore have wavefunctions, say $|\psi\rangle$ and $|\chi\rangle$, that are eigenfunctions of the energy operator H:

$$H|\psi\rangle = v|\psi\rangle, \qquad H|\chi\rangle = v|\chi\rangle.$$

Here, the energy eigenvalue v is the same in each case, because the mass is simply displaced from one location to the other. (We ignore the energy of the photon in this, the lump itself being supposed to have enormously greater mass-energy, and also any energy involvement in the device that moves the lump.) It is clear from this that there is a complete degeneracy for linear superpositions of $|\psi\rangle$ and $|\chi\rangle$, and any linear combination

$$\lambda|\psi\rangle + \mu|\chi\rangle$$

(with λ and μ complex constants) will also be an eigenstate of H with the same eigenvalue v. Each such linear combination will also be a stationary state, with the same energy eigenvalue v. Thus, any one of these combinations would be just as stable as the original two, and must therefore persist unchanged for all time.

We shall shortly consider the delicate theoretical issues that arise when one considers the (superposed) *gravitational fields* of each of the two instances of the lump also to be involved in the superposition. In accordance with this, it will be appropriate to be rather careful about the kinds of issues that become relevant when quantum theory and general relativity are considered together.

3. PRELIMINARY CONSIDERATIONS

In particular, there is a point of subtlety which needs to be addressed, even before we consider the details of any gravitational effects. We take note of the fact that, in the absence of any spatial inhomogeneity in the background potentials (gravitational or otherwise), there is nothing in the intrinsic nature of one lump location that allows us to distinguish it from any other lump location. Thus, we might choose to adopt the standpoint that there is really no physical difference at all between the various states of location of the lump and, accordingly, take the view that all these seemingly different states of the lump are actually all the same state! Indeed, this would be the normal standpoint of (quantum) general relativity. Each space-time geometry arising from of each separate lump location would be identical. The principle of general covariance forbids us to assign a meaningful (coordinate) label to each individual point, and there is no coordinate-independent way of saying that the lump occupies a different location in each of the two configurations under consideration.

However, this is not the standard attitude in ordinary quantum mechanics. If the various position states of a single quantum particle were considered to be all the same, then we would not be able to construct all the many different wave functions for that particle which are, after all, simply superpositions of states in which the particle occupies different locations. In standard quantum mechanics, all the different particle locations correspond to different quantum states whose superpositions can be independently involved.

It is worth while, in this context, to consider how standard quantum mechanics treats *composite* particles. Here one may choose to factor out by this translational freedom. For example, the wavefunction of a hydrogen atom, in an eigenstate of the energy and angular momentum operators, is still completely degenerate with respect to the spatial location of its centre of mass. In considering the quantum mechanics of a hydrogen atom, the normal procedure is (implicitly or otherwise) to factor out by the degrees of freedom that specify the location of the mass centre (or, as a simplifying approximation, merely the location of its proton), and consider the quantum mechanics only of what remains: essentially the location of the orbiting electron, this being what is fixed by the energy and angular momentum.

In the present context, with the two superposed lump locations, we are trying to examine precisely the translational degrees of freedom that had been "factored out" in the case of the hydrogen atom. In fact there is something of a physical inconsistency in simply considering our lump to be displaced from one location to another in the two states under superposition. There is a conservation law which requires that the mass centre remain fixed in space (or move uniformly in a straight line). This would have been violated unless the lump displacement is compensated by the displacement of some other massive object in the opposite direction. We shall consider that there is such another object, and that it is enormously more massive than is the lump itself. Let us call this other object "the Earth". To move the lump from one location to the other, we simply allow the Earth to move by a very tiny amount, so as to allow the mass centre to remain fixed; and since the Earth is so very much more massive than the lump, we can consider that in practice the Earth does not move at all.

The presence of the Earth in these considerations allows us to circumvent the problem that we had previously considered. For the Earth (assumed to have some large but static finite irregular shape) serves to establish a "frame of reference" against which the motion of the lump can be considered to be taking place. The space-time geometries corresponding to all the various lump locations are now all different from one another, and "general covariance" does not prevent us from considering these various states to be all distinct.

At this point, another complication appears to arise because the Earth's gravitational field has to be taken into account. In fact, we shall not be much concerned with the Earth's field. It will be the differences in the gravitational fields in the various *lump* locations that will have essential relevance. In fact, there is no problem, in the considerations which follow, in allowing the Earth's gravitational field to be involved also. But we may prefer to avoid the complication of the Earth's field, if we choose, by supposing that our "experiment" is to be set up within a spherical cavity situated at the centre of the Earth. Then the Earth's actual field could be eliminated completely. Yet the Earth would still serve to establish a "frame of reference" against which the motion of the lump could be considered to be taking place.

On the other hand, we shall prefer to consider that the lumps are actually sitting on the surface of the Earth. The Earth's (effectively) constant gravitational field does not have any significant influence on the considerations of relevance here. There is no problem about having the lump in a stationary state on the surface of the Earth, provided that some appropriate upward forces are introduced on the lump, to allow it to remain at rest with respect to the Earth.

If we were to consider a situation in which the two superposed lump positions are at different heights from the ground, then we would have to consider the effects of the energy expended in raising the lump in the Earth's field. This would introduce an additional factor, but would not seriously affect the situation that we are concerned with here. If desired, all these considerations can be evaded if we regard our lump movements to be taking place entirely within our spherical cavity at the centre of the Earth. However, it will not be necessary to pass to this extreme idealization for the considerations of this paper.

Let us now try to consider how the gravitational field of the lump itself affects our superposed state. Each lump location is accompanied by the static gravitational field produced by the lump in that location. We must envisage that the superposed state is now an entangled one,

$$\lambda|\psi\rangle|G_\psi\rangle + \mu|\chi\rangle|G_\chi\rangle,$$

where $|G_\psi\rangle$ and $|G_\chi\rangle$ are the quantum states of the gravitational fields of the lump locations corresponding to $|\psi\rangle$ and $|\chi\rangle$, respectively. We might choose to think of $|G_\psi\rangle$ and $|G_\chi\rangle$ as *coherent states*, if that is the appropriate description. In any case, whatever is to be meant by the quantum state of a stationary gravitational field — including all the internal degrees of freedom of the field — would be supposed to be incorporated into $|G_\psi\rangle$ or $|G_\chi\rangle$. But it does not greatly matter, for our present purposes, what the "correct" quantum-mechanical descriptions of macroscopic gravitational fields actually are. We must suppose, however, that whatever this description is, it closely accords in its physical interpretation, for a *single* unsuperposed lump, with the classical gravitational field of that lump according to the descriptions of Einstein's general relativity.

In fact, the "entangled" nature of the superposed state will not be of importance for us here. We simply take the two states under superposition to be

$$|\Psi\rangle = |\psi\rangle|G_\psi\rangle \qquad \text{and} \qquad |\mathrm{X}\rangle = |\chi\rangle|G_\chi\rangle.$$

The essential point is that each of the two states concerned must involve a reasonably well-defined (stationary) space-time geometry, where these two space-time geometries differ significantly from each other. We must now raise the question: is this *superposed* state still a stationary state?

We have to consider carefully what a "stationary state" means in a context such as this. In a stationary space-time, we have a well-defined concept of "stationary" for a quantum state in that background, because there is a Killing vector T in the space-time that generates the time-translations. Regarding T as a differential operator (the "$\partial/\partial t$" for the space-time), we simply ask for the quantum states that are eigenstates of T, and these will be the stationary states, i.e. states with well-defined energy values. In each case the energy value of the state in question, $|\Psi\rangle$, would be (essentially) the eigenvalue E_Ψ of T corresponding to that state:

$$T|\Psi\rangle = -i\hbar E_\Psi|\Psi\rangle.$$

However, for the superposed state we are considering here we have a serious problem. For we do not now have a specific space-time, but a superposition of two slightly differing space-times. How are we to regard such a "superposition of space-times"? Is there an operator that we can use to describe "time-translation" in such a superposed space-time? Such an operator would be needed so that we can identify the "stationary states" as its eigenvectors, these being the states with definite energy. It will be shown that there is a fundamental difficulty with these concepts, and that the notion of time-translation operator is *essentially* ill defined. Moreover, it will be possible to define a clear-cut measure of the *degree* of this ill-definedness for such a superposed state. Accordingly, there is, in particular, an essential uncertainty or "fuzziness" in the very concept of energy for such a state, and the degree of this uncertainty can be estimated in a clear-cut way. This is consistent with the view that such a superposed state is *unstable*, and the lifetime of the state will be given by $\hbar$ divided by this measure of energy uncertainty, in accordance with the way in which Heisenberg's uncertainty principle is employed in the theory of unstable particles, where the particle's lifetime is related to its mass-energy uncertainty.

How are we to regard a quantum superposition of two space-times? It is not sufficient to take a completely formal attitude to such matters, as is common in discussions of quantum gravity. According to the sorts of procedure that are often adopted in quantum gravity, the superposition of different space-times is indeed treated in a very formal way, in terms of complex functions on the space of 3-geometries (or 4-geometries), for example, where there is no pretence at a pointwise identification of the different geometries under superposition. A difficulty with such formal procedures arises, however, if we attempt to discuss the physics that takes place *within* such a formal superposition of spaces, as is the case with the type of situation under consideration here (cf. also Ref. 1).

Indeed, in the case of the two minutely differing space-times that occur in our situation with the two superposed lump locations, there would be no obvious way to register the fact that the lump is actually in a different place, in each of the two configurations under superposition, unless there *is* some sort of (approximate) identification. (Of course, in this identification, since the lump itself is supposed to be "moved", corresponding points of the lump are not identified, but the corresponding points of the Earth are, since the Earth is not considered to have significantly moved.)

Nevertheless, it is clear from the principles of general relativity that it is not appropriate, in general, to make a precise identification between points of one space-time and corresponding points of the other. The *gauge* *freedom* of general relativity — as reflected in "the principle of general covariance" (or, equivalently, "diffeomorphism invariance") — is precisely the freedom that forbids a meaningful precise labelling of individual points in a space-time. Accordingly there is generally no precise meaningful pointwise identification between different space-times. (In special cases, it may be possible to circumvent this problem; see Refs. 1,2, where issues of this nature are examined, and interesting effects are anticipated, in relation, in particular, to the quantization of a cosmic string.) In the general case, all that we can expect will be some kind of *approximate* pointwise identification.

5. THE SEMICLASSICAL APPROACH

One possible way to address this difficulty might be to adopt a "semiclassical approximation" to quantum gravity, according to which it would be the *expectation value* of the (quantum) energy-momentum tensor that serves to specify the right-hand side of Einstein's equations

$$R_{ab} - \frac{R}{2}\, g_{ab} = -8\pi G T_{ab}\,.$$

In such an approximation the superposition of our pair of states in which one lump is in two spatially displaced locations would have a gravitational field which is merely the average field of the fields of each lump individually.

In fact, the semiclassical approximation is not really physically consistent. In particular, it allows superluminary communication (see Ref. 29); and, in a certain interpretation, it is grossly inconsistent with observation [26]. These difficulties might be avoided if the semiclassical interpretation can be combined with some scheme of gravitationally induced state-vector reduction, as is argued by Pearle and Squires (Ref. 29; cf. also Ref. 23).

From the point of view of the present paper, there is an additional difficulty with the semiclassical approximation. The semiclassical description would provide a space-time containing the mass distributions of two spatially displaced lumps with vacuum ($R_{ab} = 0$) between them. This gravitational field — of a pair of spatially displaced lumps — is not really the same as a linear superposition of the two fields, each describing one of the two lump locations individually. The gravitational interaction effects between the pair of lumps would have to be taken into account, according to the non-linear effects of general relativity. Assume that the actual lump, in its two displaced locations, sits on a smooth horizontal table ("the Earth"). The "semiclassical" state would represent a pair of spatially separated massive lumps, and would not actually possess an exact Killing vector which could play the role of T (="$\partial/\partial t$") because the "semiclassical" lumps would fall towards each other along the table in accordance with their gravitational attractions, leading to a non-static space-time geometry.

For reasons such as these, it is proposed not to follow a semiclassical description here. Instead, we shall follow the route of supposing that there is some kind of *approximate sense* in which the two superposed space-times can be pointwise identified. Without some kind of (approximate) identification between the space-times, we do not seem even to be able to express the fact that the quantum state of the lump is a superposition of two distinct locations, in which the degree of displacement between the two lump locations is reasonably well defined.

6. APPROXIMATE SPACE-TIME POINT IDENTIFICATION

The basic principles of general relativity — as encompassed in the term "the principle of general covariance" (and also "principle of equivalence") — tell us that there is no natural way to identify the points of one space-time with corresponding points of another. Consider our quantum superposition between two different space-times (here, the fields of two alternative lump locations). If we are to attempt to make a pointwise identification between these two space-times, we can do this in a way that would be only *approximately* meaningful. Let us try to obtain some measure of this degree of approximation.

In order to proceed to a reasonably explicit expression of this measure, it will be helpful to make the assumption that a Newtonian approximation to the gravitational fields of each lump location is adequate. Indeed, in any plausible practical situation in which a quantum superposition of lump locations might have to be considered, this would certainly be the case. Thus, there will be spatial sections of the two space-times in question which are Euclidean 3-spaces, and each space-time will possess a well-defined time coordinate whose constant values define these Euclidean spatial sections. We shall suppose that the time coordinates for each space-time can be naturally identified with each other, so we have a single time parameter t common to the two space-times. This would seem reasonable for the Newtonian situation under consideration, although there would be essential subtleties arising when the two gravitational fields are treated according to full general relativity.

It should be made clear, however, that passing to the Newtonian limit does not remove the difficulties that the principle of general covariance — or the principle of equivalence — presents in relation to the quantum superposition of gravitational fields. As may be recalled, there is a space-time formulation of Newtonian gravitational theory originally provided by Cartan [7], and further studied independently by Friedrichs [14] and Trautman [40], in which the Newtonian version of the equivalence principle is directly incorporated into the geometrical description. (See Ref. 12 for an up-to-date account.) The Cartan geometric formulation of Newtonian gravitational theory is indeed the appropriate one for taking into account the subtleties of (what remains of) the principle of general covariance and the principle of equivalence. (For example, the Newton-Cartan space-time of a *constant* non-zero Newtonian gravitational field has an identical geometry to that of the zero gravitational field, but it differs from a non-uniform Newtonian field which produces tidal effects.) It has been pointed out by Christian [8] that the Newton–Cartan framework provides a valuable setting for exploring some of the fundamental problems of unifying quantum theory with gravitational theory without, at this stage, the more severe difficulties of general relativity having to be faced. Christian argues that this framework indeed sheds important light on the role of gravity in the measurement problem.

As it turns out, the criterion for quantum state reduction that we shall be led to here is independent of the value of the speed of light c. It will thus have a well-defined non-trivial Newtonian limit, and it can be expressed within the Newton–Cartan framework. However, although it will be valuable to bear this framework in mind here, we shall not actually use its specific mathematical details.

The essential point about superposing a pair of Newton–Cartan space-times is that whereas we are allowing that the time-coordinate t can be identified in the two space-times — and so there is a canonical correspondence between the various space sections of one space-time with those of the other — there is no canonical way of identifying the individual *points* of a section of one space-time with corresponding points of the other. It is this lack of a definite pointwise identification between the spatial sections of the two space-times which will lead us to an essential ill-definedness of the notion of time-translation — and therefore of the notion of stationarity — for the quantum-superposed state.

The reader might wonder why it is *time-translation* that should encounter problems, when, in our Newtonian context, there is no problem with the time-coordinate t. However, this is how it should be; for "time-translation" is something which is represented by the operator "$\partial/\partial t$", and the meaning of this operator is really concerned more with the choice of the *remaining* variables x, y, z, (those parameters to be held fixed in the definition of $\partial/\partial t$) than it is with t itself. An uncertainty in the *pointwise* identification of the spatial sections of one space-time with the spatial sections of the other will indeed show up in an uncertainty in the definition of time-translation. Suppose that we use the coordinates x, y, z, and t for the points of one space-time, and x', y', z', and t' for the points of the other (where, of course, $t = t'$); then we have

$$\frac{\partial}{\partial t'} = \frac{\partial}{\partial t} + \frac{\partial x}{\partial t'}\frac{\partial}{\partial x} + \frac{\partial y}{\partial t'}\frac{\partial}{\partial y} + \frac{\partial z}{\partial t'}\frac{\partial}{\partial z}$$

$$= \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla$$

where $\mathbf{v}$ is the spatial velocity, as described with respect to the unprimed coordinate system, of a point fixed in the primed coordinate system. (In the (x, y, z, t)-system, the 3-vector $\mathbf{v}$ has components $(\partial x/\partial t', \partial y/\partial t', \partial z/\partial t')$, i.e. $(-\partial x'/\partial t, -\partial y'/\partial t, -\partial z'/\partial t)$, and the operator ∇ has components $(\partial/\partial x, \partial/\partial y, \partial/\partial z)$.) If we were to attempt to identify the point with coordinates (x, y, z, t) in one of the space-times under consideration, with the point with coordinates (x', y', z', t') in the other, then we should encounter an incompatibility between their notions of time-evolution unless $\mathbf{v}$ vanishes everywhere.

Of course, in this Newtonian limit, it *is* possible to arrange that $\mathbf{v} = 0$ everywhere, simply by taking ordinary static, non-rotating Newtonian/Cartesian coordinates for the two space-times, related by a constant spatial displacement between them. But this would be to go against the spirit of what is entailed, in the present context, by Einstein's principle of general covariance. It is the special nature of the Newtonian limit that provides us with flat Euclidean spatial sections whose local motions are determined directly by what happens at infinity. In this Newtonian limit, the "coordinate freedom" can indeed be eliminated if we nail things down at infinity, because in Newton–Cartan geometry, there is exact spatial rigidity. However, within the more general context of curved (and not necessarily stationary) space-times, this is not at all appropriate. Our physical expectations would be for the criteria characterizing the nature of a localized quantum superposition to depend on reasonably local criteria. The principles of general relativity are antagonistic towards the idea of identifying individual local points in a precise way, in terms of the situation at infinity.

Similar remarks also apply to certain proposals within the general programme of quantizing general relativity proper, according to which the time-translation operator would not be taken to have local relevance, but to refer merely to the symmetries at spatial infinity — it being supposed that only asymptotically flat space-times are to be considered in quantum superposition. It seems, however, for situations such as those under consideration here where we are concerned with a reasonably local problem, that it is quite inappropriate to define the notion of time-translation merely in terms of what is going on at spatial infinity. We do not know, after all, what the geometry of the actual universe is on a very large scale (spatially asymptotically spherical or hyperbolic, for example, either of which would lead to different notions of "time-translation" from that which is normally considered). It is most unlikely that, for a reasonably local quantum problem such as this, Nature should really "care" what is going on at infinity.

We shall suppose instead that, in the particular Newtonian quantum superposition under consideration, it is appropriate to demand an approximate spatial identification between the two space-times, where the degree of approximation is governed by local (or quasi-local) considerations in the vicinity of the region where the identification occurs. In accordance with the principle of equivalence, it is the notion of free fall which is locally defined, so the most natural local identification between a local region of one space-time and a corresponding local region of the other would be that in which the free falls (i.e. space-time geodesics) agree. However, in the superposition under consideration, there is no way to make the spatial identifications so that the free falls agree everywhere throughout the space-times. The best that one can do is to try to minimize the amount of the *difference* between free fall motions.

How are we to express this difference mathematically? (Let us assume that the two quantum amplitudes assigned to the two superposed states are about the same size. Then we may take it that neither space-time's geometry dominates the other.) Let $\mathbf{f}$ and $\mathbf{f}'$ be the acceleration 3-vectors of the free-fall motions in the respective space-times, at some identified point (where the accelerations can be taken with respect to the appropriate local identified coordinates). In fact, $\mathbf{f}$ and $\mathbf{f}'$ will be the Newtonian gravitational force-per-unit-test-mass, at that point, in each space-time. Let us take the scalar quantity

$$(\mathbf{f}-\mathbf{f}')^2 = (\mathbf{f}-\mathbf{f}')\cdot(\mathbf{f}-\mathbf{f}')$$

as the measure of incompatibility of the identification — and, accordingly, of the "uncertainty" involved in this identification. It may be noted that if we apply an *acceleration* to the common coordinates to the two space-times, while keeping the actual point identification unaltered, then the difference $\mathbf{f}-\mathbf{f}'$ does not change. In fact, the expression $(\mathbf{f}-\mathbf{f}')^2$

is coordinate-independent (assuming that the "dot" $\cdot$ refers to the actual 3-dimensional spatial metric).

For the total measure of incompatibility (or "uncertainty") Δ at a particular instant, given by a particular t-value, it is proposed to integrate this quantity (with respect to the spatial 3-volume element d^3x) over the t = constant 3-space:

$$\begin{aligned}\Delta &= \int (\mathbf{f} - \mathbf{f}')^2 d^3x \\ &= \int (\nabla\Phi - \nabla\Phi')^2 d^3x \\ &= \int (\nabla\Phi - \nabla\Phi') \cdot (\nabla\Phi - \nabla\Phi') d^3x \\ &= -\int (\Phi - \Phi')\,(\nabla^2\Phi - \nabla^2\Phi') d^3x\end{aligned}$$

where $\mathbf{f} = -\nabla\Phi$ and $\mathbf{f}' = -\nabla\Phi'$, Φ and Φ' being the respective gravitational potentials for the two space-times. By Poisson's formula

$$\nabla^2\Phi = -4\pi G\rho,$$

we obtain

$$\Delta = 4\pi G \int (\Phi - \Phi')\,(\rho - \rho') d^3x,$$

where ρ and ρ' are the respective mass densities. Using the integral formula

$$\Phi(x) = -\int \rho(y)\,/\,|x - y|\, d^3y,$$

we get

$$\Delta = -4\pi G \iint (\rho(x) - \rho'(x))\,(\rho(y) - \rho'(y))\,/\,|x - y|\, d^3x\, d^3y,$$

which is basically just the gravitational self-energy of the *difference* between the mass distributions of each of the two lump locations.

How does this relate to the uncertainty in the time-translation Killing vector referred to at the beginning of this section? That there should *be* some relationship follows from general considerations, but the exact form that this relationship should take seems to depend upon the specific model that is used to describe the uncertainty of space-time identification. (As long ago as 1966, Károlyházy provided some fairly closely related considerations, cf. Ref. 20 and also Refs. 21,22.) To understand what is involved in the present context, we consider three successive time slices, given by $t = \tau$, $t = \tau + \Delta t$, and $t = \tau + 2\Delta t$. We imagine that there is some uncertainty in the identification between the two space-times under superposition at each of these three times. The measure of relative "acceleration uncertainty" that we have just been considering, and which seems to be forced upon us from consideration of the principles of equivalence and general covariance, has to do with the difference between the error in spatial identification at $t = \tau + \Delta t$ and the average of those at $t = \tau$ and at $t = \tau + 2\Delta t$. The overall uncertainty in the quantity $\mathbf{v} \cdot \nabla$, which appears as the uncertainty in the time-translation Killing vector for the superposed space-times is a "velocity uncertainty", and it has to do with the difference between the error in spatial identification at $t = \tau + \Delta t$ and that at $t = \tau$ (or else between the error at $t = \tau + 2\Delta t$ and that at $t = \tau + \Delta t$). The precise relationship between the "acceleration uncertainty" and "velocity uncertainty" seems to depend upon the way that this uncertainty is actually modelled. The direct "position uncertainty" does not feature in these considerations; it is the time-evolution of the error in spatial identification that has importance for us here.

A further ingredient of possible importance is an error in the identification in the actual time coordinate t for the two space-times. We have ignored this complication here, but it clearly has relevance in full general relativity. Perhaps it should also be taken into account in the Newtonian limit.

In view of these various complicating issues, no attempt will be made to formulate a definitive statement of the precise measure of uncertainty that is to be assigned to the "superposed Killing vector" and to the corresponding notion of "stationarity" for the superposed space-time. However, it is strongly indicated by the above considerations that the quantity Δ, as defined above, gives a very plausible (though provisional) estimate of this uncertainty. Hence, we can use Δ, or some simple multiple of this quantity, as a measure of the fundamental energy uncertainty "E_Δ" of the superposed state. Accordingly, the superposed state would not be exactly stationary, but it is consistent with the above considerations that it should have a lifetime of the general order of $\hbar/E_\Delta$.

It is reassuring that basically the same expression is seen to arise from certain other considerations. For example, it was pointed out in [35,36] that the gravitational self-energy involved in a quantum superposition of a pair of differing space-time geometries should involve an essential uncertainty, owing to the fact that even in classical general relativity there is difficulty with the energy concept for gravity. (There is no local expression for gravitational energy.) Moreover, considerations of the symplectic structure of linearized gravity [34] lead to something very similar, and so also does the earlier model of Diósi [11], although in these papers the suggestion is that the expression

$$G \iint \rho(x)\rho'(y) \,/\, |x-y| \, d^3x \, d^3y,$$

which is the *gravitational interaction energy*, should be the relevant measure of the required "energy uncertainty" E_Δ. In fact (apart from the factor 4π), it does not make any difference which expression is used, provided that — as is the case here — the two individual states under superposition each have the same gravitational self-energy. (A situation for which the difference in the expressions might be important would be in a cloud chamber, where the two states in superposition might be "droplet forming" and "droplet not forming". In this case the gravitational self-energy of the droplet would contribute a difference between the two expressions.)

7. FURTHER CONSIDERATIONS

It is clear that the above arguments provide only a very preliminary analysis of the difficulties involved in the notion of "stationarity" for a quantum superposition of differing stationary space-times. However, even just these preliminary considerations seem to suggest that such a superposition would be unstable, and that a very plausible expression for the order of magnitude of the lifetime of such a superposed state (where the two relevant amplitudes are roughly equal) is $\hbar/E_\Delta$, where E_Δ is the gravitational self-energy of the *difference* between the two mass distributions involved.

In any case, it should be clear from what has been said above that *conventional quantum theory* provides no clear answer (in the absence of a satisfactory theory of quantum gravity) to the problem of the stability of a quantum superposition of two different gravitating states. We do not need to appeal to the contentious conceptual issues inherent in the measurement problem of quantum mechanics for some motivation for believing that such superpositions should be unstable.

In the above considerations, we have restricted attention to cases where the superposition involves just *two* different states, each of which individually has a well-defined space-time geometry, and where the two amplitudes assigned to each of these two constituent states are of about the same size. This does not tell us, for example, what to do about the wavefunction of a single isolated proton, as it spreads throughout space in accordance with the Schrödinger equation. In the GRW scheme [16] there would be a small probability for this state to reduce spontaneously to a more localized state. The present considerations, however, are not adequate to provide us with an expectation as to whether (or how frequently) such a spontaneous state reduction should occur, nor do they tell us what kind of state should be the result of such a spontaneous state reduction. We might think that the "natural" states for the proton, having "well-defined space-time geometries", are those for which the proton is reasonably localized, say to a region of roughly its Compton wavelength. But if we adopt such a view, we are driven to consider that most wavefunctions are superpositions of a great many of these natural states, and with widely differing amplitudes. A more detailed theory is clearly needed if such questions are to be addressed adequately.

Despite these uncertainties, it is still possible, in many different circumstances, to estimate the expected order of magnitude of the rate of gravitationally induced state-vector reduction according to this scheme. For a single proton, we may expect that a superposed state of two separated spatial locations will decay to one or the other location in something of the order of a few million years. For a water speck 10^{-5}cm in radius, the timescale would be about an hour or so; for a speck 10^{-3}cm in radius, something like a millionth of a second. These results indeed seem reasonable, and if confirmed would supply a very plausible solution to the quantum measurement problem, but for the moment they appear to be rather beyond what can be experimentally tested.

It should be emphasized that none of the considerations of the present paper give any clear indication of the mathematical nature of the theory that would be required to incorporate a plausible gravitationally induced spontaneous state-vector reduction. In all probability, such mathematical considerations would have to come from quite other directions. Indeed, this author's own expectations are that no fully satisfactory theory will be forthcoming until there is a revolution in the description of quantum phenomena that is of as great a magnitude as that which Einstein introduced (in the description of gravitational phenomena) with his general theory of relativity.

ACKNOWLEDGEMENTS

I am grateful to Abhay Ashtekar, Jeeva Anandan, Joy Christian, Ted Newman, and Lee Smolin and others for some very helpful discussions.

I am also grateful to NSF for support under research contract PHY 93-96246.

REFERENCES

1. Anandan, J. (1994). *Gen. Rel. Grav.* **26**, 125.
2. Anandan, J. (1995). "Gravitational phase operator and cosmic strings". To appear in *Phys. Rev. D.*
3. Bell, J. S. (1990). *Physics World* **3**, 33.
4. Bialynicki-Birula, I., and Mycielski, J. (1976). *Ann. Phys. (NY)* **100**, 62.
5. Bohm, D. (1952). *Phys. Rev.* **85**, 166.
6. Bohm, D., and Hiley, B. (1994). *The Undivided Universe* (Routledge, London).
7. Cartan, É. (1923,1924). "Sur les variétés à connexion affine et la théorie de la relativité generalisée", *Ann. École Norm. Sup.* **40**, 325, **41**, 1.
8. Christian, J. (1995). "Definite events in Newton–Cartan quantum gravity" Oxford University preprint.
9. de Broglie, L (1956). *Tentative d'Interpretation Causale et Nonlineaire de la Mechanique Ondulatoire* (Gauthier-Villars, Paris).
10. DeWitt, B. S., and Graham, R. D. (eds.) (1973). *The Many Worlds Interpretation of Quantum Mechanics* (Princeton University Press, Princeton).
11. Diósi, L. (1989). *Phys. Rev. A***40**, 1165.
12. Ehlers, J. (1991). In *Classical Mechanics and Relativity: Relationship and Consistency (Int. Conf. in memory of Carlo Cataneo, Elba 1989)*, G. Ferrarese, ed. (Monographs and Textbooks in Physical Science, Lecture Notes 20, Bibliopolis, Naples).
13. Everett, H. (1957). *Rev. Mod. Phys.* **29**, 454.
14. Friedrichs, K. (1927). *Math. Ann.* **98**, 566.
15. Gell-Mann, M., and Hartle, J. B. (1993). *Phys. Rev. D***47**, 3345.
16. Ghirardi, G. C., Rimini, A. and Weber, T. (1986). *Phys. Rev. D***34**, 470.
17. Ghirardi, G. C., Grassi, R., and Rimini, A. (1990). *Phys. Rev. A***42**, 1057.
18. Griffiths, R. (1984). *J. Stat. Phys.* **36**, 219.
19. Haag, R. (1992). *Local Quantum Physics: Fields, Particles, Algebras* (Springer-Verlag, Berlin).
20. Károlyházy, F. (1966). *Nuovo Cimento A***42**, 390.
21. Károlyházy, F. (1974). *Magyar Fizikai Polyoirat* **12**, 24.
22. Károlyházy, F., Frenkel, A., and Lukács, B. (1986). In *Quantum Concepts in Space and Time*, R. Penrose and C. J. Isham, eds. (Oxford University Press, Oxford), p.109.
23. Kibble, T. W. B. (1981). In *Quantum Gravity 2: a Second Oxford Symposium*, C. J. Isham, R. Penrose and D. W. Sciama, eds. (Oxford University Press, Oxford), p.63.
24. Komar, A. B. (1969). *Int. J. Theor. Phys.* **2**, 157.
25. Omnès, R. (1992). *Rev. Mod. Phys.* **64**, 339.
26. Page, D. N., and Geilker, C. D. (1981). *Phys. Rev. Lett.* **47**, 979.
27. Pearle, P. (1985). In *Quantum Concepts in Space and Time*, R. Penrose and C. J. Isham, eds. (Oxford University Press, Oxford), p.84.
28. Pearle, P. (1989). *Phys. Rev. A***39**, 2277.
29. Pearle, P., and Squires, E. J. (1995). "Gravity, energy conservation and parameter values in collapse models". Durham University preprint DTP/95/13.
30. Penrose, R. (1981). In *Quantum Gravity 2: A Second Oxford Symposium*, C. J. Isham, R. Penrose and D. W. Sciama, eds. (Oxford University Press, Oxford), p.244.

31. Penrose, R. (1986). In *Quantum Concepts in Space and Time*, R. Penrose and C. J. Isham, eds. (Oxford University Press, Oxford), p.129.
32. Penrose, R. (1987). In *300 Years of Gravity*, S. W. Hawking and W. Israel, eds. (Cambridge University Press, Cambridge), p.17.
33. Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press, Oxford).
34. Penrose, R. (1993). In *General Relativity and Gravitation 13. Part 1: Plenary Lectures*, R. J. Gleiser, C. N. Kozameh and O. M. Moreschi, eds. (IOPP, Bristol/Philadelphia), p.179.
35. Penrose, R. (1994). In *Fundamental Aspects of Quantum Theory*, J. Anandan and J. L. Safko, eds. (World Scientific, Singapore), p.238.
36. Penrose, R. (1994). *Shadows of the Mind; An Approach to the Missing Science of Consciousness* (Oxford University Press, Oxford).
37. Percival, I. C. (1995). *Proc. R. Soc. Lond. A***451**, 503.
38. Squires, E. (1990). *Phys. Lett. A***145**, 67.
39. Squires, E. (1992). *Found. Phys. Lett.* **5**, 279.
40. Trautman, A. (1965). In *Lectures on General Relativity (Brandeis 1964 Summer Inst. on Theoretical Physics)*, vol. I, by A. Trautman, F. A. E. Pirani and H. Bondi (Prentice-Hall, Englewood Cliffs, NJ), p.7.
41. Weinberg, S. (1989). *Phys. Rev. Lett.* **62**, 485.
42. Wigner E. P. (1961). In *The Scientist Speculates*, I. J. Good, ed. (Heinemann, London); reprinted in E. Wigner (1967). *Symmetries and Reflections* (Indiana University Press, Bloomington) and in 1983 in *Quantum Theory and Measurement*, J. A. Wheeler and W. H. Zurek, eds. (Princeton University Press, Princeton).
43. Zurek, W. H. (1991). *Physics Today* **44**, 36.

253

TWISTOR THEORY, THE EINSTEIN EQUATIONS, AND QUANTUM MECHANICS

Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, World Scientific, 1996

802 Twistor theory, the Einstein equations, and quantum mechanics. *Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation* (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, 1996

TWISTOR THEORY, THE EINSTEIN EQUATIONS, AND QUANTUM MECHANICS

Roger Penrose
Mathematical Institute
University of Oxford, England

1. The question of quantized space-time and quantum gravity

The normal procedures of quantum gravity involve leaving the principles of quantum mechanics (QT) intact, whilst attempting to modify general relativity (GR) - or some other classical space-time theory - in accordance with those principles. Yet GR is tested overall to an even greater accuracy (about 1 part in 10^{14} as implied by the precision of the Hulse Taylor observations of the binary pulsar PSR 1913+16 over a period of 20 years), than is QT (to 1 part in 10^{11} or 10^{12} as follows from the agreement between theory and experiment in quantum electrodynamics).Thus, my own viewpoint is to seek a new theory which modifies and units both GR and QT, and for which both emerge as limiting cases of this theory. *Twistor theory* represents an attempt in that direction.

Before embarking on a description of twistor theory, however, it will be appropriate to consider what I regard as the main problem of present-day quantum theory, namely how the quantum micro-world relates to the classical macro-world. There would appear to be one set of laws for the micro-world, governed by complex numbers (where the amplitudes have their moduli squared to become real-number probabilities). The fundamental difficulties is well illustrated by Schrödinger's famous cat.

Here I shall consider that a photon is emitted at a source and is put into a superposition of a transmitted and reflected state by a half-silvered mirror. The transmitted part of the photon's wave function causes a gun to fire which kills a cat, so that according to the standard unitarity procedures of quantum state evolution (Schrödinger's equation), the cat ought to be put into a superposition of death and life at the same time - with complex coefficients! (see Fig.1).

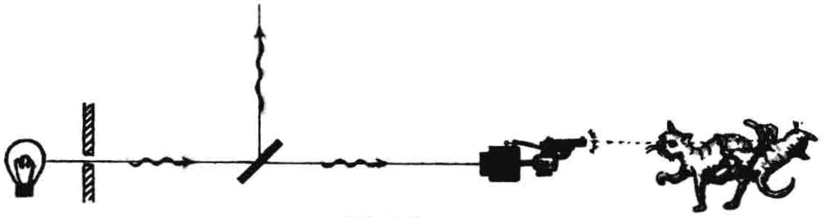
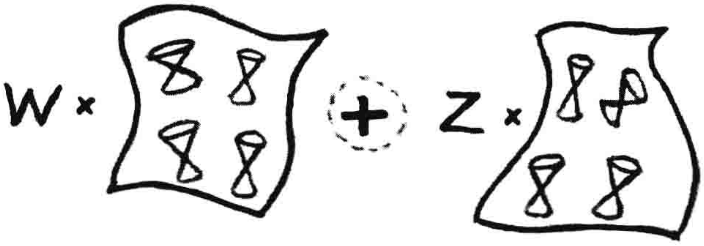


Fig.1

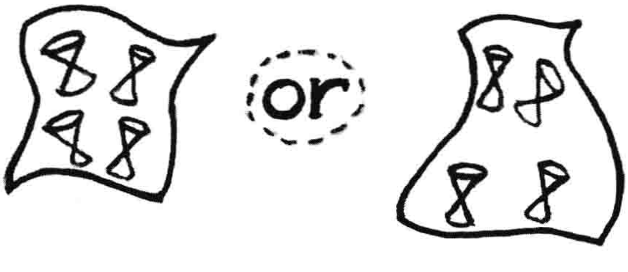
Schrödinger's cat
a half-silvered mirror splits a phothon's state into a
superposition of two states, one of which kills the cat.
Is the cat in a superposition of dead and alive?

But perhaps something goes wrong with superposition principle when it applies to significantly differing space-time geometries.

The live cat and the dead one will indeed be accompained by space-time geometries which slightly differ, so a superposition



of two space-times might actually become



(with relative probabilities $|w|^2 : |z|^2$).

I shall call such a putative actual physical process Objective Reduction. Accordingly, we need a new **OR** theory!

Of course, from this point of view, there is no requirement that the macroscopic objects under superposition need be anything so complicated as a cat. Let us suppose that the transmitted photon simply moves a lump of material from one location to another (see Fig.2)

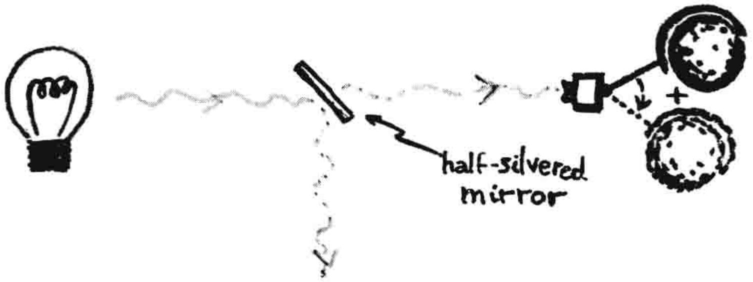


Fig.2

An inanimate version of Schrödinger's cat, with a lump of material being put into a superposition of two locations. For how long can this persist?

I am not going to attempt to provide such a theory here. But we can at least pose the following question: At what stage does a superpositiom (+) of the two lump positions become either one lump position or the other? My suggestion is as follows. Consider the energy E that it would take to displace one instance of the lump away from the gravitational field of the other. I am proposing that the superposed state is unstable (like that of an unstable nucleus), and collapses (decays) to one or the other individual state, in an approximate time $T \sim \hbar/E$

(Also Kávolyhásy, Ghirardi, Rimini, Weber, Grassi, Pearle, and particularly Diósi, have made similar suggestion; see refs.)

Stability of a superposition of a pair of lump locations

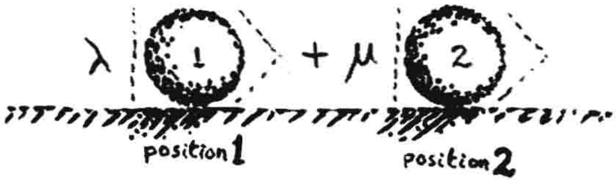


Fig.3

I want to give you a new argument supporting the idea that such superpositions of macroscopically displaced mass distributions are unstable. Let $|\psi_1\rangle$ and $|\psi_2\rangle$ be the quantum states for two alternative spatially displaced stationary lump locations. Ignoring the gravitational fields produced by the lump itself, in each location, we have a timelike Killing vector **T** for background static space-time, which plays the role of the "$\partial/\partial t$" in the Schrödinger equation. For each individual state, we have

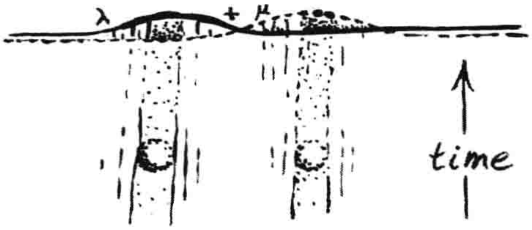
$$i\hbar\mathbf{T}|\psi_1\rangle = F|\psi_1\rangle, \qquad i\hbar\mathbf{T}|\psi_2\rangle = F|\psi_2\rangle,$$

where F is its energy, whence for the superposed state,

$$i\hbar\mathbf{T}\left(\lambda|\psi_1\rangle + \mu|\psi_2\rangle\right) = F\left(\lambda|\psi_1\rangle + \mu|\psi_2\rangle\right).$$

We have a complete degeneracy: all superpositions of $|\psi_1\rangle$ and $|\psi_2\rangle$ are just as stationary as are $|\psi_1\rangle$ and $|\psi_2\rangle$ individually.

Now, bring in the gravitational field of the lump. Are all the superposed states still stationary, each with well-defined energy (F)?



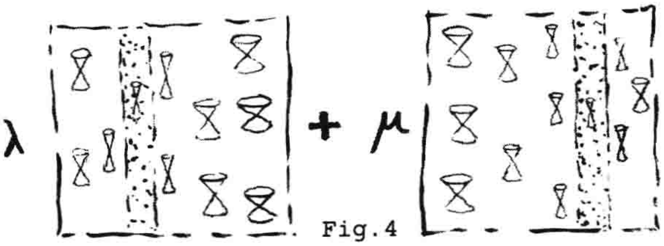


Fig.4

Superposition of differing space-times?

We have to consider superpositions of actual differing *space-time geometries*, since this is how the gravitational fields of the lump in its two alternative locations must be described. Quantum gravity usually treats such a superposition very formally, but we need to do physics *in* such a superposed space-time. It is hard to see how to make sense of such an idea - even to express the fact that the lump occupies a different location in the superposed space-times - without some kind of pointwise identification between the space-times. However the principle of general covariance is antagonistic to such a proposal and, moreover, such an identification would lead to a *superposed causality* (see Fig.5), which does not make sense. Accordingly, I

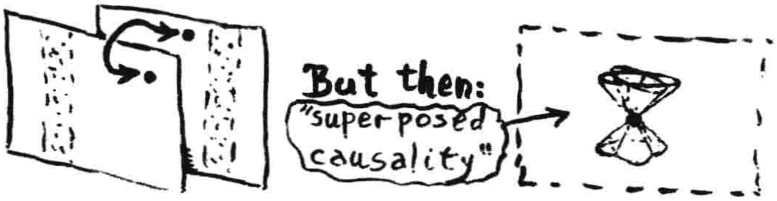


Fig.5

Try to identify points? - at least approximately shall consider that there is some approximate notion of pointwise identification between the two space-times under superposition, where there is an imbuilt "uncertainty involved in the identification. This will lead to an ill-definedness in the notion of a Killing vector for the superposed space-times, and hence to an ill-definedness in the very notion of stationarity.

We can take as the "measure of ill - definedness" of the Killing vector, the local expression

$$\overset{1}{g}_{00}/\overset{2}{g}_{00} \quad \text{or} \quad \phi_1-\phi_2$$

(where $\sqrt{g_{00}} \approx 1+2\phi$, with ϕ = Newtonian potential; $G=1$)

or, rather, as its spatial variation. In the Newtonian limit, we can express this variation as

$$\vec{\nabla}(\phi_1-\phi_2)$$

and then we can take as the total measure of ill-definedness, the expression

$$\int \vec{\nabla}(\phi_1-\phi_2)\cdot\vec{\nabla}(\phi_1-\phi_2)\,d^3x$$

$$= -\int(\phi_1-\phi_2)\,\nabla^2(\phi_1-\phi_2)\,d^3x$$

$$=4\pi\int(\phi_1-\phi_2)\,(\rho_1-\rho_2)\,d^3x$$

$$=-4\pi\iint\frac{(\rho_1(x)-\rho_2(x))(\rho_1(y)-\rho_2(y))}{|x-y|}\,d^3x\,d^3y$$

where ρ is the mean density. This is just the gravitational self-energy of the difference mass distribution (i.e. of one mass distribution subtracted pointwise from the other).

In the case of a pair mass distributions in which one is just a rigid displacement of the other, this gives just the same energy

E as we had before, for the superposed displaced lump locations. (In that case E was considered to be the gravitational interaction energy, which was specifically Diósi's proposal. The two agree for mere rigid displacements, but it is not altogether clear which is a better measure, in the general case). The lack of definiteness in the notion of stationarity corresponds to this energy E, so we anticipate a lifetime of the superposed state of the order of $\hbar E^{-1}$ as asserted earlier.

For an isolated system, the lump movement itself is the only mass diaplacement which need be considered. In this case we can easily estimate $\hbar E^{-1}$ in simple cases. For a nucleon, considered as a solid sphere of about a fermi's radius in a superposition of two separate spatial locations, the state would jump to one location on the other in a timescale of a few million years. Of course, this need not be a very realistic model, but it tells us that we can ignore **OR** effects for a single nucleon which is just as well, since neutron interferometry is a confirmed fact). For a water speck of 10^{-3}cm radius, however, a decay time for the superposed state would be about a millionth of second. If of radius 10^{-4}cm (micron), we expect decay of the superposition in about 1/20 second; if of radius 10^{-5}cm, then it might take a few hours (Fig.6).

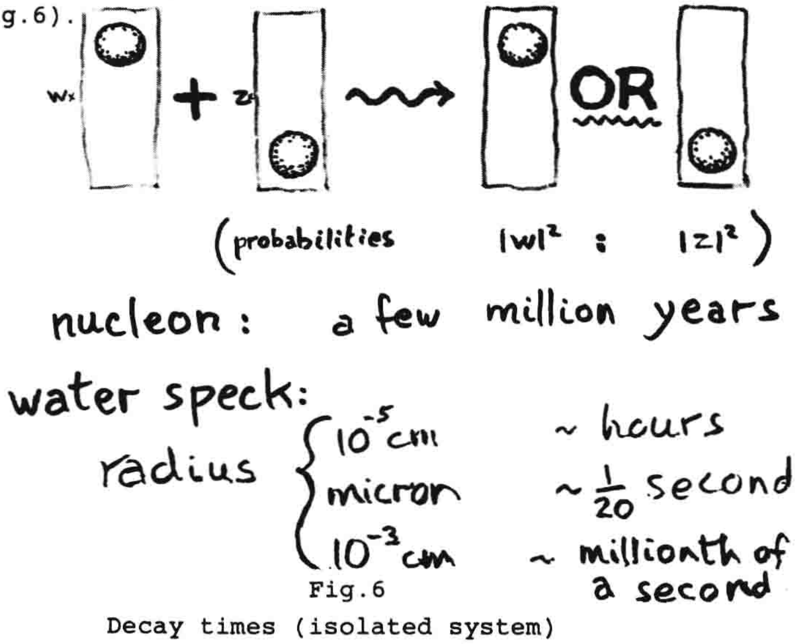

w×
+
z×
OR
(probabilities |w|² : |z|²)
nucleon: a few million years
water speck:
radius
10⁻⁵ cm ~ hours
micron ~ 1/20 second
10⁻³ cm ~ millionth of a second

Fig.6

Decay times (isolated system)

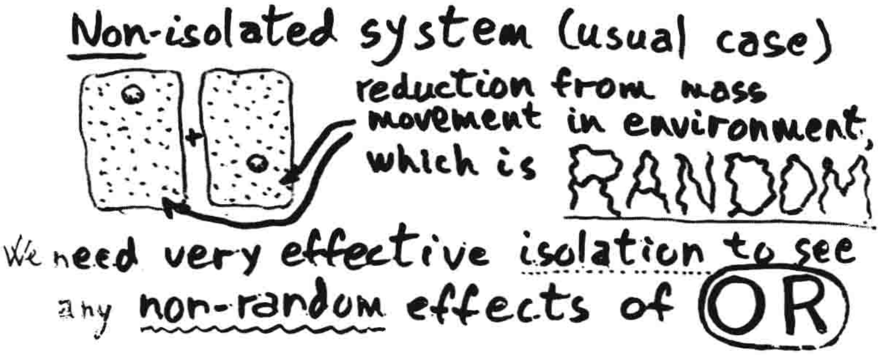

Non-isolated system (usual case)
+
reduction from mass movement in environment, which is RANDOM
We need very effective isolation to see any non-random effects of OR

Fig.7

The effect of the environment on state-reduction

These times do not take into account the fact that, in most physical situations, the system under consideration is not isolated, so there may be a good deal of mass movement in the environment accompanying the movement of the lump. The quantum state will be an entangled one, and the major part of the contribution to E is then likely to come from the movement of the particles, in the environment, which are entangled with the state of the lump. Since the environment normally consists largely of randomly moving particles, their effect on the **OR** process would be an essentially random one, but where reduction would take place much more rapidly than would have been the case without the entangled environment. (see Fi.7). This leads, in effect, to nothing noticeably different from what is expected in the normal treatments of "reduction" due to environmental entanglement. It seems that the effects of an **OR** theory, consistent with the ideas being expressed here, would be very hard (though not impossible) to detect experimentally.

There would be many very difficult problems to be faced by any detailed **OR** theory. One of the most severe of these comes about because of the quantum entanglements that arise in EPR (Einstein-Podolsky-Rosen) -type situations. The most famous example of an experimentally verified EPR effect occurs with the experiment performed by Aspect and colleagues in 1982. Here there is a photon source which emits pairs of photons in opposite directions, and the polarization plane of each photon is measured, the two measurements being some 12 metres apart. The selections of directions in which the two polarizations are measured is not made until the photons are in full flight. Yet, the pair of photons behaves as a single unit right up to the time of measurement. According to a famous theorem of J.S.Bell, there is no way to explain the joint probabilities of the pair of polarization measurements that quantum theory predicts (and Aspect's experiment confirms) in terms of a picture in which the two photons are treated as separate independent objects.

The difficulty of this with regard to relativity is that with an entangled system (such as the pair of photons just described), the measurement of one member of the system appears instantaneously to affect the state of the other. Such a description would not be consistent with the principles of relativity (see Fig.8). This suggests to me that a radically different approach to space-time description will be needed if a fully objective picture of entangled quantum states is to be obtained. This is what twistor theory strives and achieves. In the following sections I shall try to indicate where this theory stands at present.

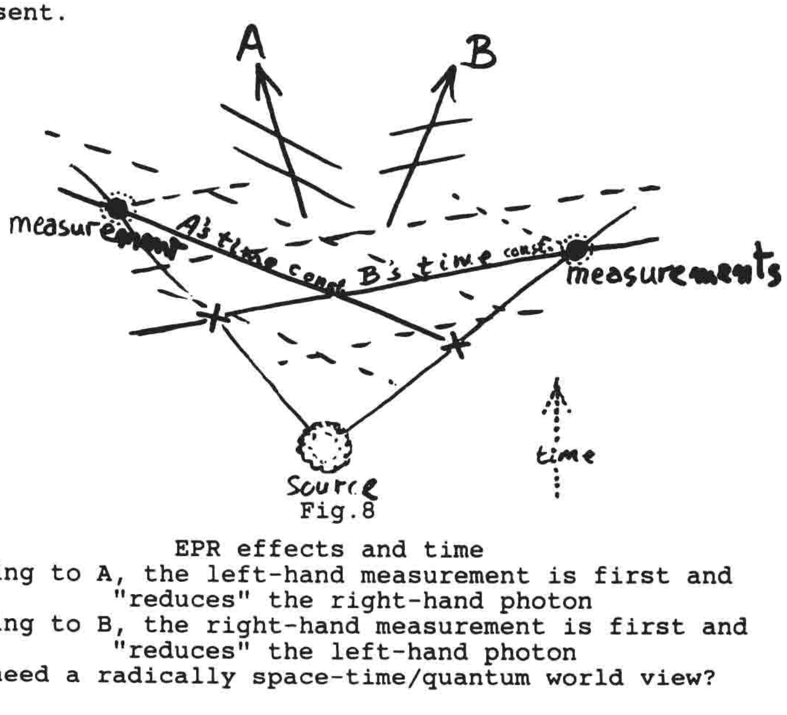


Fig.8

EPR effects and time

According to A, the left-hand measurement is first and "reduces" the right-hand photon

According to B, the right-hand measurement is first and "reduces" the left-hand photon

Do we need a radically space-time/quantum world view?

2. Motivations for twistor theory

The theory of twistors is partially motivated by a desire for a non-local approach to space-time physics, where space-time points (events) are secondary, the twistors themselves being primary in this picture. Other motivations come from the complex number structure of QM.

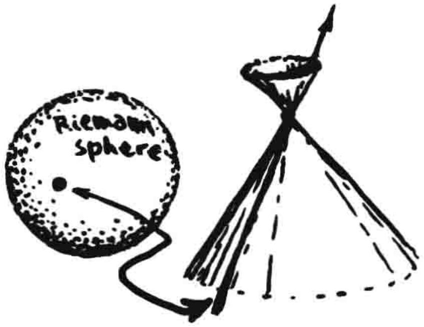


Fig.9
Twistors as light rays
The light rays through any space-time point constitute a Riemann sphere.

Although the complex numbers (amplitudes) that appear in quantum mechanics are usually treated rather abstractly, there is actually a fairly close relationship between these complex numbers and space-time geometry, even though this relationship is not always very manifest and is not very widely appreciated. The most obvious place where this relationship can be seen occurs with an ordinary spin 1/2·particle, such as electron, proton, or neutron. If we consider the two basic states to be "spin up" and "spin-down, described by the respective state vectors $|\uparrow\rangle$ and $|\downarrow\rangle$, the general state of spin

$$u\ |\uparrow\rangle\ +\ v\ |\downarrow\rangle$$

turns out to be eigenstate of a spin measurement in some other direction, the various possible spatial directions being representable as points on a Riemann sphere (complex plane together with a point at infinity). The complex number $z = v/u$ can be represented on this sphere by stereographic projection to the complex plane (see Fig.10). This shows up the close relationship between complex amplitudes and spatial geometry in the case of a spin 1/2 particle.

812 Twistor theory, the Einstein equations, and quantum mechanics. *Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation* (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, 1996

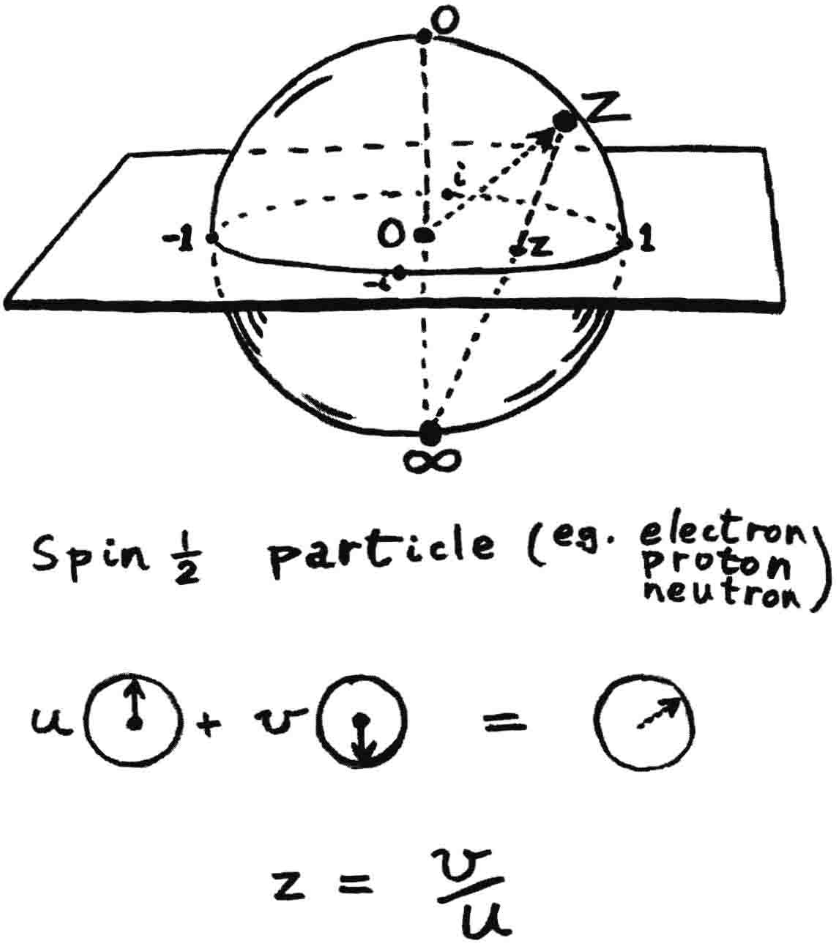

0
Z
i
-1
0
z
1
-i
∞
Spin ½ particle (eg. electron proton neutron)
u + v =
z = v/u

Fig.10

Spin 1/2: Stereographic projection of the Riemann sphere. It represents the complex ratio z = v/u, indicating a close relationship between spatial geometry and complex quantum amplitudes.

This might be thought to be a rather special property of spin 1/2. However, it turns out that there is a corresponding geometrical relationship for higher spin also. In Fig.11, the Majorana description for the general spin state, of spin n/2 is indicated. Here, there are n (unordered) points on the Riemann sphere representing the general state. They define the n rotes of a polynomial whose coefficients correspond to the various amplitudes for the eigenstates of spin in the upward direction. These eigenstates are the states with definite m-value - those states distinguished by a Stern-Gerlach apparatus with field oriented upwards. They have Majorana descriptions for which there are (1/2)n + m points all at the north pole and (1/2) - m points all at the south pole. See Fig.11.

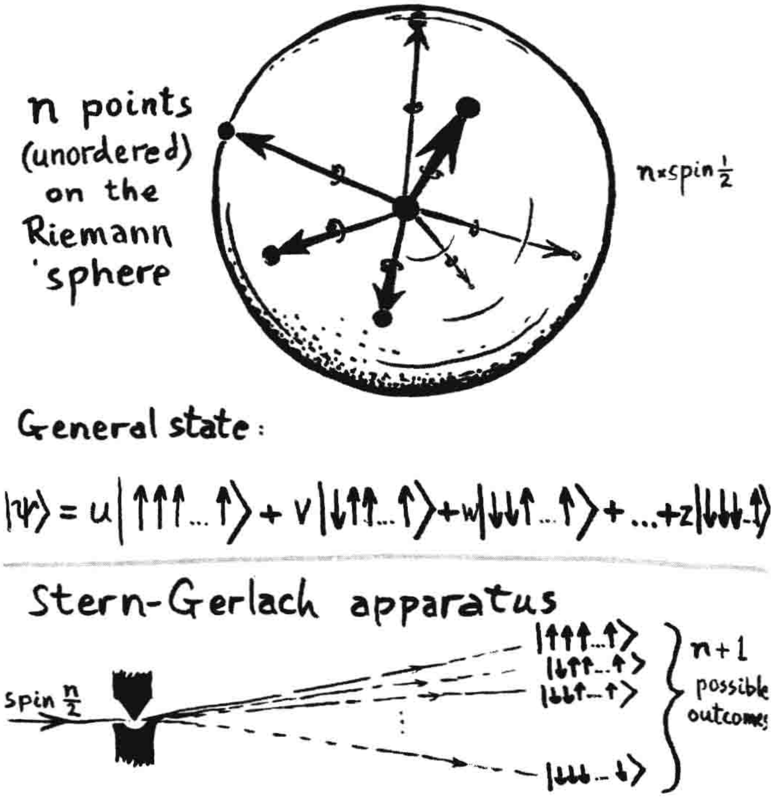

n points (unordered) on the Riemann sphere
n×spin ½
General state:
|ψ⟩ = u|↑↑↑…↑⟩ + v|↓↑↑…↑⟩+w|↓↓↑…↑⟩+…+z|↓↓↓…↓⟩
Stern-Gerlach apparatus
spin n/2
|↑↑↑…↑⟩
|↓↑↑…↑⟩
|↓↓↑…↑⟩
|↓↓↓…↓⟩
n+1 possible outcomes

Fig.11
Majorana description for higher spin states: spin n/2

The Riemann sphere also has a role to play in relativity, as the celestial sphere of an observer. The reason that this is appropriate is that the transformations that take the celestial sphere of an observer who passes close by the first observer and looks out at the same sky, but transformed by abberation effects or by a rotation of frame of reference, are precisely the (holomorphic) transformations of the Riemann sphere. In other words, the action of the (restricted) Lorentz group on the celestial sphere is the same as the action of global complex-analytic maps of the Riemann sphere to itself.

The geometric nature of these transformation follows from the properties of stereographic projection: angles are preserved and circles map to circles (see Fig.12).

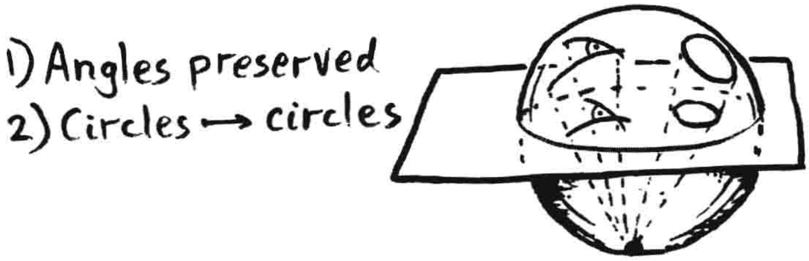


Fig.12
Properties of stereographic projection

This property also holds for the celestial sphere (Fig.13). The general transformation of the sphere preserving angles (also circles) and orientation is given by

$$u \Longrightarrow u' = (\alpha u + \beta)/(\gamma u + \beta),$$

where u is a (stereographic) complex coordinate on the sphere - which projects to an ordinary complex coordinate on the plane. This is called a Möbius transformation. It is a remarkable fact that the most fundamental group in physics - the non-reflective Lorentz group - is indeed the Möbius group of the Riemann sphere - the symmetry group of the simplest complex manifold (namely the Riemann sphere). This fact, which relates complex geometry to physics at its deepest level of space-time structure, is at the

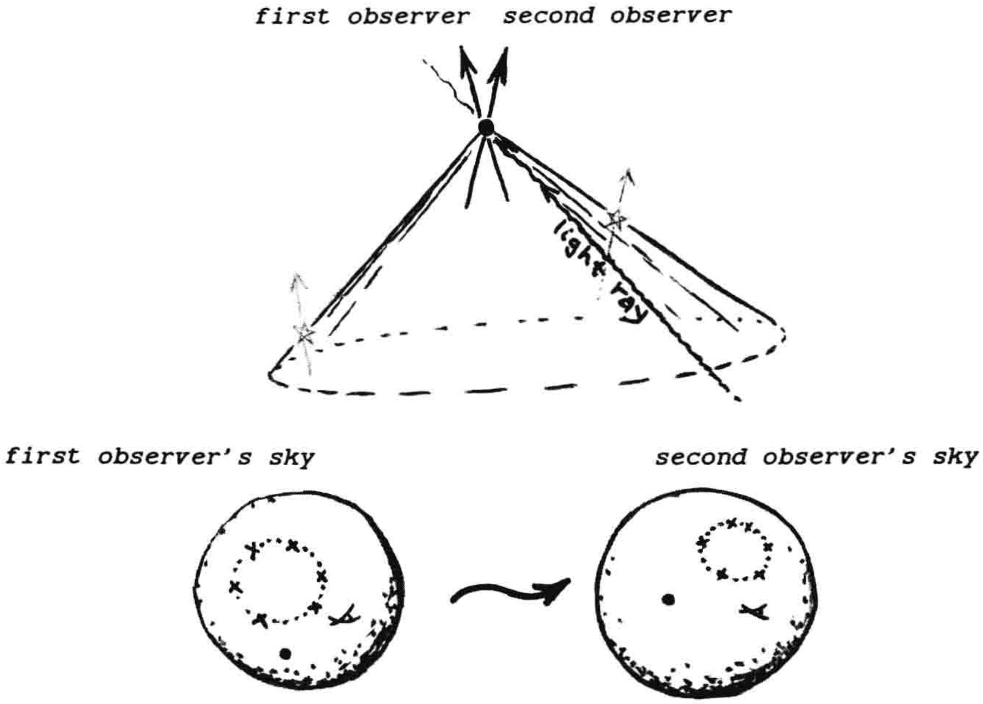


Fig.13

The conformal and circle-preserving property of Lorentz transformations

basis of the 2-spinor formalism. Carrying this philosophy further leads us into twistor theory.

3. Twistor geometry

Twistor geometry provides a different way of studying space-time structure from that provided by the usual formalism. Instead of space-time points (events) being the fundamental entities, we think of entire light rays as being more fundamental. Because of the Riemann-sphere structure of the celestial sphere, a point R of the space-time - as represented by the family of light rays through that point - corresponds to a Riemann sphere in light-ray space. As a first approximation to twistor theory, we may think of light-ray space as twistor space, so a point R in space-time is represented by a Riemann sphere R in twistor space whereas a point Z in twistor space is represented by a light ray Z in space-time (Fig.14).

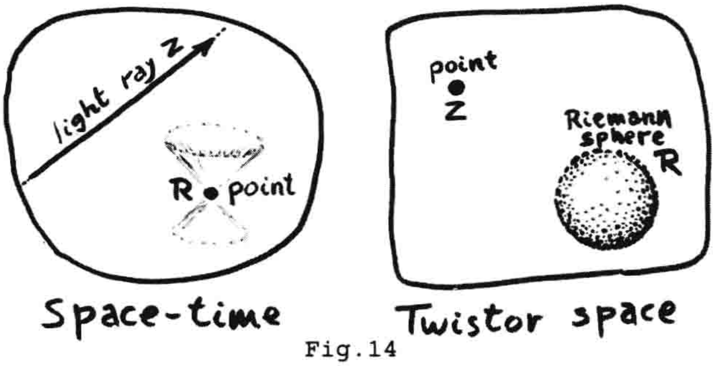

light ray Z
R point
point Z
Riemann sphere R
Space-time
Twistor space

Fig.14
A first approximation to twistor theory

The general philosophy is that this twistor correspondence is to be fundamentally complex, in addition to being basically non-local. Thus, we should like twistor space to turn out to be a complex manifold of some kind. A basic problem arises, however, bacause light-ray space is 5-dimensional, whereas all complex manifolds must have even numbers of real dimensions. It turns out that for Minkowski space-time $\mathbb{M}$, there is an elegant solution to this problem. Light-ray space is 5-real-dimensional sub-manifold of a 6-real-dimensional manifold $\mathbb{PT}$ (called projective twistor space) which has a natural structure as a complex projective 3-space ($\mathbb{CP}^3$). The Riemann spheres corresponding to points in $\mathbb{M}$ are projective lines ($\mathbb{CP}^1$) in $\mathbb{PN}$ (Fig.15).

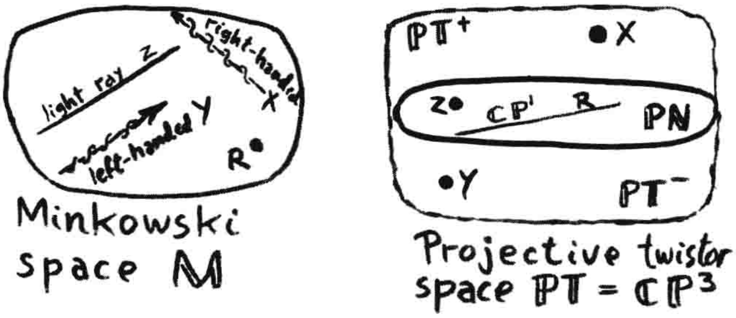

light ray Z
right-handed X
left-handed Y
R
Minkowski space $\mathbb{M}$
$\mathbb{PT}^+$
X
Z
$\mathbb{CP}^1$
R
$\mathbb{PN}$
Y
$\mathbb{PT}^-$
Projective twistor space $\mathbb{PT} = \mathbb{CP}^3$

Fig.15
The basic twistor correspondence

To see how this correspondence comes about, take ordinary Minkowski coordinates for $\mathbb{M}$. A point R has space-time coordinates (r^0, r^1, r^2, r^3) while Z has twistor coordinates (Z^0, Z^1, Z^2, Z^3). The basic correspondence is expressed in the condition of incidence

$$\begin{pmatrix} Z^0 \\ Z^1 \end{pmatrix} = \frac{i}{\sqrt{2}} \begin{pmatrix} r^0+r^3 & r^1+ir^2 \\ r^1-ir^2 & r^0-r^3 \end{pmatrix} \begin{pmatrix} Z^2 \\ Z^3 \end{pmatrix}.$$

Geometrically, incidence expresses:

in $\mathbb{M}$	in $\mathbb{PT}$
point R lies on light ray Z	point Z lies on projective line R.

The space $\mathbb{PT}$ is defined by projective twistors, i.e. Z^α up to proportionality, i.e. by the ratios $Z^0:Z^1:Z^2:Z^3$. The twistors themselves define the non-projective vector space π.

In order to proceed further with twistor theory, it is convenient to introduce the 2-spinor notation. Here, each 4-tensor index a,b,c,.... is replaced by a pair of 2-spinor indices, one unprimed and one primed: a = AA', b = BB', c = CC',.... (using abstract indices). In standard coordinates, the vector

$$V^a = V^{AA'}$$

has components

$$\begin{pmatrix} V^{00'} & V^{01'} \\ V^{10'} & V^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} V^0+V^3 & V^1+iV^2 \\ V^1-iV^2 & V^0-V^3 \end{pmatrix}$$

whence

$$\det V^{AA'} = \tfrac{1}{2}\left[(V^0)^2-(V^1)^2-(V^2)^2-(V^3)^2\right] = \tfrac{1}{2} g_{ab}V^aV^b$$
$$= \tfrac{1}{2}\,\varepsilon_{AB}\,\varepsilon_{A'B'}\,V^{AA'}V^{BB'}$$

The quantities ϵ_{AB} , $\epsilon_{A'B'}$ are skew, having components which are standard Leci Civita symbols (0,1;-1,0). The spinor form of the metric tensor is then

$$g_{ab} = \varepsilon_{AB}\,\varepsilon_{A'B'}$$

the ϵ-quantities are used for lowering and raising spinor indices according to

$$\xi^{A} = \varepsilon^{AB}\xi_{B} \qquad \xi_{A} = -\varepsilon_{AB}\xi^{B}$$

Primed and unprimed indices are interchanged under complex conjugation. A single-index spinor ξ^A has (up to sign) a real-space-time representation as a null flag, i.e. as a null vector $x^a = \xi^A\bar{\xi}^{A'}$ together with an oriented null 2-plane element containing that null direction (see Fig.16).

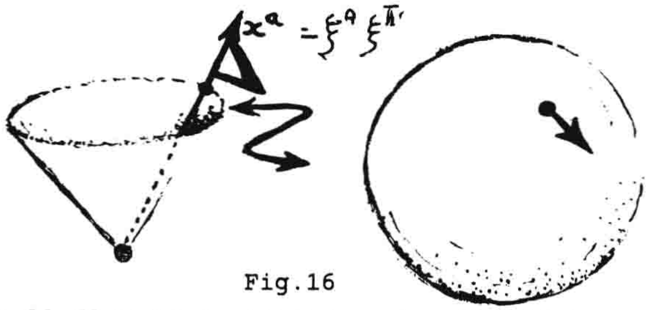


Fig.16
The null flag representation of a 2-spinor

Let O be an origin-point in $\mathbb{M}$. We can express the incidence relation as

$$\omega^{A} = i\Gamma^{AA'}\pi_{A'} ,$$

where $\Gamma^{a} = \Gamma^{AA'}$ is the position vector of some point R on a light ray Z, and where the four complex components of the twistor Z^{α} are the two components of each of the 2-spinors ω^A and $\pi_{A'}$, respectively. We write

$$Z^{\alpha} = (\omega^{A}, \pi_{A'})$$

and call ω^A and $\pi_{A'}$ the spinor parts of the twistor Z^{α}. The geometrical significance of ω^A and $\pi_{A'}$ (up to proportionality) is that the flagpole of $\pi_{A'}$ points along Z, whereas the flagpole of ω^A points at two light ray Z from the origin O (see Fig.17).

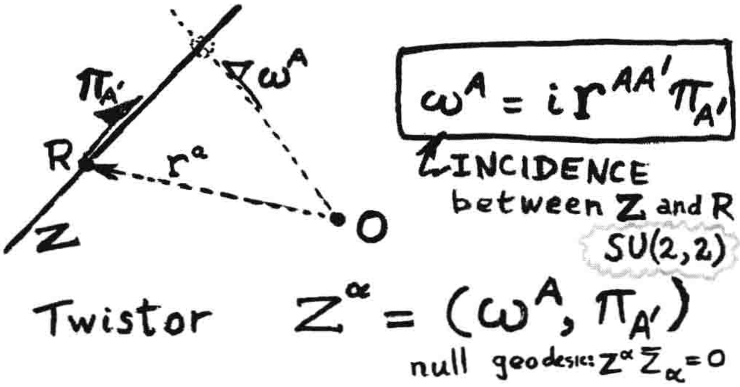

$\pi_{A'}$
ω^A
R
r^a
Z
O
$\omega^A = i r^{AA'} \pi_{A'}$
INCIDENCE between Z and R
SU(2,2)
Twistor $Z^\alpha = (\omega^A, \pi_{A'})$
null geodesic: $Z^\alpha \bar{Z}_\alpha = 0$

Fig.17

The spinor parts of a null twistor

The complex conjugate $\bar{Z}$ of a twistor Z^α is a dual twistor with spinor parts defined by

$$\bar{Z} = (\bar{\pi}_A\ ,\ \bar{\omega}^{A'})$$

For a null ray, we must have

$$Z\,\bar{Z}_\alpha = 0, \text{ i.e. } \omega^A\bar{\pi}_A + \pi_{A'}\bar{\omega}^{A'} = 0\ ,$$

this being the equation of ℙ𝕅, as depicted in Fig.15. When this equation does not hold, there are no real points R incident with the twistor Z^α. However, there will be complex points R (points of ℂ𝕄, the complexification of 𝕄) which are incident with Z^α, i.e. whose position vectors $r^{AA'}$ (not required to be Hermitian) satisfy the incidence relation $\omega^A = ir^{AA'}\pi_{A'}$ (so long as $\pi_{A'} \neq 0$, otherwise the points R would lie at infinity). For fixed Z^α, the complex points R incident with Z^α constitute a complex totally null plane (so of four real dimensions) called an α-plane. The tangent vectors to an α-plane are all null vectors, having the spinor form

$$\lambda^A\pi^{A'}$$

where $\pi^{A'}$ is fixed and λ^A varies. There is also another kind of totally null plane in ℂ𝕄, called a β-plane, whose tangent vectors have this same form but where now it is λ^A which is held fixed and $\pi^{A'}$ which varies. The points S of a β-plane are those incident with a dual twistor $W_\alpha = (\lambda_A\ ,\ \mu^{A'})$, the incidence relation of S with W_α being expressed as

$$\mu^{A'} = -i\lambda_A s^{AA'}$$

where $s^a = s^{AA'}$ is the position vector (with respect to O) of S. See Fig.18 for the geometry of incidence of R with Z^α and S with W_α.

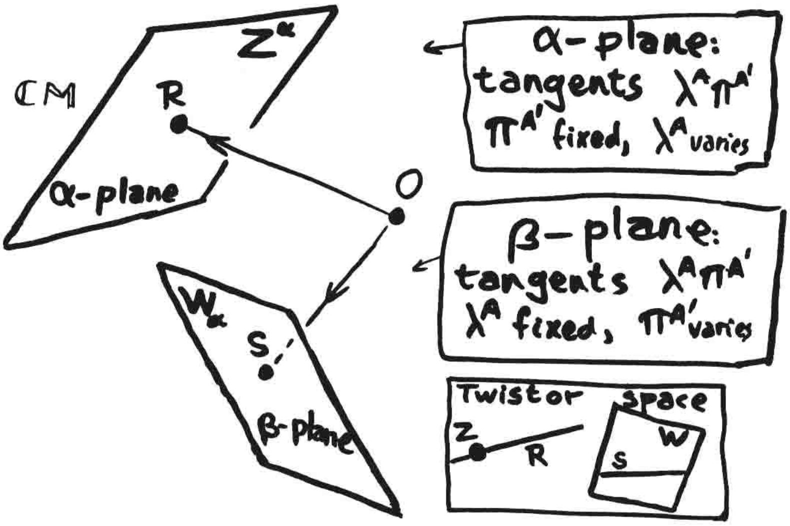

ℂM
Z^α
R
α-plane
O
W_α
S
β-plane
α-plane: tangents $\lambda^A\pi^{A'}$ $\pi^{A'}$ fixed, λ^A varies
β-plane: tangents $\lambda^A\pi^{A'}$ λ^A fixed, $\pi^{A'}$ varies
Twistor space
Z
R
W
S

Fig.18
Complex picture

We can also express an incidence of Z^α with W_α as

$$Z^\alpha W_\alpha = 0$$

which means, geometrically, that the α-plane Z and the β-plane W have a point in common. When this happens we also have a whole line of intersection of the α-plane with the β-plane (Fig.19).

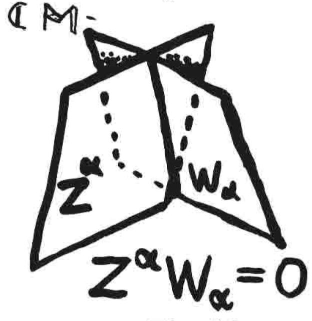

ℂM
Z^α
W_α
$Z^\alpha W_\alpha = 0$

Fig.19
Incidence between an α-plane and a β-plane. They then meet in a complex null geodesic

Such a line is a *complex* light ray in $\mathbb{CM}$. Any complex light ray has a unique α-plane through it, and also a unique β-plane. (When the light ray is real, tha α-plane and the β-plane are complex conjugates of one another.)

Complexified Minkowski space, when compactified by the addition of appropriate elements at infinity can be thought of as a complex quadric in $\mathbb{CP}^5$. This is the *Klein quadric* representing the projective twistor space $\mathbb{PT}$. See Fig.20.

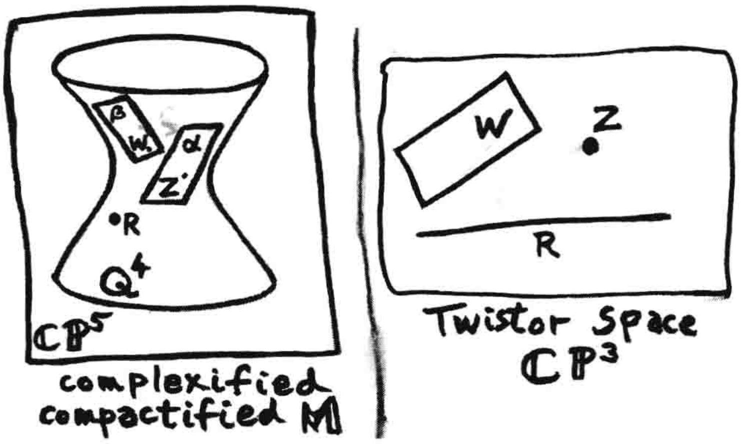


Fig.20
Klein correspondence

Basically, the kind of twistor description that I have been giving up to this point is what might be referred to as a *Mk.1 twistor*, given by a light ray (null geodesic) in Minkowski space $\mathbb{M}$, or by an α-plane in $\mathbb{CM}$. We shall shortly be coming to a *Mk.2 twistor*, which describes a spinning massless particle in terms of the 4-momentum and angular momentum structure for massless particle in $\mathbb{M}$. There is also the notion of a *Mk.3 twistor*, which is a helicity raising operator for massless fields in $\mathbb{M}$. There are partial generalizations of Mk.1 twistors to curved space-time $\mathcal{M}$: hypersurface twistors, asymptotic twistors, local twistors, 2-surface twistors, ambitwistors etc., which I shall not discuss here, but later on I shall say something about Mk.4 twistors, which may be thought of as charges for massless helicity 3/2 fields. These are consistent in Ricci-flat general curved space-times, but not fully understood, as yet.

4. Twistors, angular mementum, and helicity

A Mk.2 twistor represents the 4-momentum P_a and the 6-angular momentum M^{ab} structure of a massless particle:

$$P_a = P_{AA'} = \bar{\pi}_A \pi_{A'}$$

$$M^{ab} = M^{AA'BB'} = i\omega^{(A}\bar{\pi}^{B)}\varepsilon^{A'B'} - i\varepsilon^{AB}\bar{\omega}^{(A'}\pi^{B')}$$

(where the parentheses around indices denote symmetrization). The helicity is a real number s, given by

$$s = \tfrac{1}{2} Z^{\alpha}\bar{Z}_{\alpha} = \tfrac{1}{2}\{\omega^{A}\bar{\pi}_{A} + \pi_{A'}\bar{\omega}^{A'}\}$$

where the Pauli-Lubanski spin vector is proportional to the momentum, with proportionality factor s:

$$\tfrac{1}{2}\epsilon_{abcd}\, p^{b} M^{cd} = s\, p_a$$

This enables us to interpret twistors that are not null as "spinning photons", rather than "light rays" (see Fig.15). In terms of explicit components, these relations are

$$\begin{aligned}
p^0 &= \tfrac{1}{\sqrt{2}}(Z^3\bar{Z}_1 + Z^2\bar{Z}_0) \\
p^1 &= \tfrac{1}{\sqrt{2}}(Z^3\bar{Z}_1 - Z^2\bar{Z}_0) \\
p^2 &= \tfrac{1}{\sqrt{2}}(-Z^2\bar{Z}_1 - Z^3\bar{Z}_0) \\
p^3 &= \tfrac{i}{\sqrt{2}}(Z^2\bar{Z}_1 - Z^3\bar{Z}_0)
\end{aligned}$$

Twistor theory, the Einstein equations, and quantum mechanics. *Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation* (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, 1996

for the 4-momentum components

$$M^{01} = -M^{10} = \frac{i}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 - Z^2\bar{Z}_2 + Z^3\bar{Z}_3)$$
$$M^{02} = -M^{20} = \frac{i}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 - Z^2\bar{Z}_3 - Z^3\bar{Z}_2)$$
$$M^{03} = -M^{30} = \frac{1}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 - Z^2\bar{Z}_3 + Z^3\bar{Z}_2)$$
$$M^{12} = -M^{21} = \frac{i}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 + Z^2\bar{Z}_3 - Z^3\bar{Z}_2)$$
$$M^{32} = -M^{23} = \frac{1}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 + Z^2\bar{Z}_2 - Z^3\bar{Z}_3)$$
$$M^{13} = -M^{31} = \frac{1}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 + Z^2\bar{Z}_3 + Z^3\bar{Z}_2)$$
$$M^{00} = M^{11} = M^{22} = M^{33} = 0$$

for the 6-angular momentum components are

$$(\bar{Z}_0, \bar{Z}_1, \bar{Z}_2, \bar{Z}_3) = (\overline{Z^2}\ \overline{Z^3}, \overline{Z^0}, \overline{Z^1}).$$

For the helicity, we have

$$s = (1/2)(Z^0\bar{Z}_0 + Z^1\bar{Z}_1 + Z^2\bar{Z}_2 + Z^3\bar{Z}_3).$$

To understand the role of Mk.3 twistor we consider what happens when we move the origin O,in $\mathbb{M}$, to a new point X, and see how ω^A and $\pi_{A'}$ change, keeping the twistor Z^α fixed in its geometrical interpretation. We shall think of X as a variable field point, instead of using the fixed origin in O (see Fig.21.)

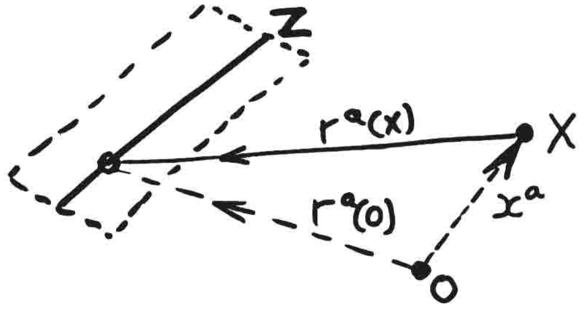


Fig.21
Making the origin a variable point

$(\omega^A, \pi_{A'})$ become fields (keeping Z fixed) which satisfy

$$\nabla^{A'(A}\omega^{B)}=0$$

which is called the twistor equation (which is conformally invariant). From this equation follows the existence of a spinor field $\pi_{A'}$ such that $\nabla_{AA'}\omega^{B}=-i\varepsilon_{A}{}^{B}\pi_{A'}$

$$\nabla_{AA'}\pi_{B'}=0$$

The twistor equation provides us with a twistor's role as a spin lovering operator. Take a massless field

$$\phi_{AB\ldots LM} = \phi_{(AB\ldots LM)}$$

subject to the field equation

$$\nabla^{AA'}\phi_{AB\ldots LM} = 0$$

then the spin-lovered field

$$\phi_{AB\ldots LM}\omega^{M} = \psi_{AB\ldots L}$$

satisfies same equation, but with one fewer indices.

5. Twistor quantum theory

Let us try to use twistor space as the framework for a quantum theory of (massless) particles. The classical state of a massless particle can be described by a single twistor; in quantum theory, complex linear superpositions of these must be considered, i.e. we have a function $f(Z^\alpha)$, where f is a complex function. However, this is not a wave function yet, because Z^α describes both position and momentum variables. We must take into account the twistor commutation rules:

$$[Z^\alpha,\bar{Z}_\beta]=\hbar\delta^\alpha_\beta,\quad [Z^\alpha,Z^\beta]=0,\quad [\bar{Z}_\alpha,\bar{Z}_\beta]=0.$$

Thus, Z^α and $\bar{Z}_\beta$ are conjugate variables, and a wave function should not involve both. A twistor wave function is a function of

Z^α independent of $\overline{Z}_\alpha$, which means it is holomorphik in Z^α. We can also consider dual twistor wave functions, which are anti--holomorphic in Z^α, i.e. holomorphic in a dual twistor W_α (= $\overline{Z}_\alpha$).

The non-commutation of Z^α with $\overline{Z}_\beta$ could lead to possible factor ordering problems in the expressions for momentum and angular momentum given earlier. In fact, the expressions for p_a and M^{ab} are actually unaffected, but we find

$$s = (1/4)\left\{Z^\alpha\,\overline{Z}_\alpha + \overline{Z}_\alpha\,Z^\alpha\right\} = (\hbar/2)\left\{-\ 2 - Z^\alpha(\partial/\partial Z^\alpha)\right\} =$$

$$= (\hbar/2)\left\{-\ 2 - \textit{homogeneity degree}\right\}$$

or, in terms of $W_\alpha = \overline{Z}_\alpha$,

$$s = (\hbar/2)\left\{2 + W_\alpha(\partial/dW_\alpha\right\} = (\hbar/2 + \textit{homogeneity degree})$$

The relationships between spin and helicity are illustrated in the table

TABLE

	$S/\hbar$	hom. deg. in Z^α	hom. deg. in W_α
graviton	+2	−6	+2
photon	+1	−4	0
anti-neut.	+1/2	−3	−1
scalar	0	−2	−2
neutrino	−1/2	−1	−3
photon	−1	0	−4
graviton	−2	+2	−6

6. Twistor contour integrals

How do we get from a twistor wavefunction f to a space-time field ϕ.... ? This is done by means of a contour integral

$$\phi_{\cdot\cdot\cdot\cdot\cdot}(r^{\cdot}) = \oint \begin{matrix}\pi_{\cdot}\pi_{\cdot}\cdots\ \pi_{\cdot}\\ \text{or}\\ \frac{\partial}{\partial\omega^{\cdot}}\frac{\partial}{\partial\omega^{\cdot}}\cdots\frac{\partial}{\partial\omega^{\cdot}}\end{matrix} f(Z^{\cdot})\ \pi_{\cdot}d\pi^{\cdot}$$

where Z is constrained to be incident with **r**, i.e. where $\omega = i\mathbf{r}\pi$, $Z = (\omega, \pi)$, the two components of ω^A being Z^0, Z^1, and the two components of $\pi_{A'}$ being Z^2, Z^3. The function f needs to be holomorphic in an appropriate region for the contour integration to give a non-trivial answer. Such a region is illustrated in Fog.22:

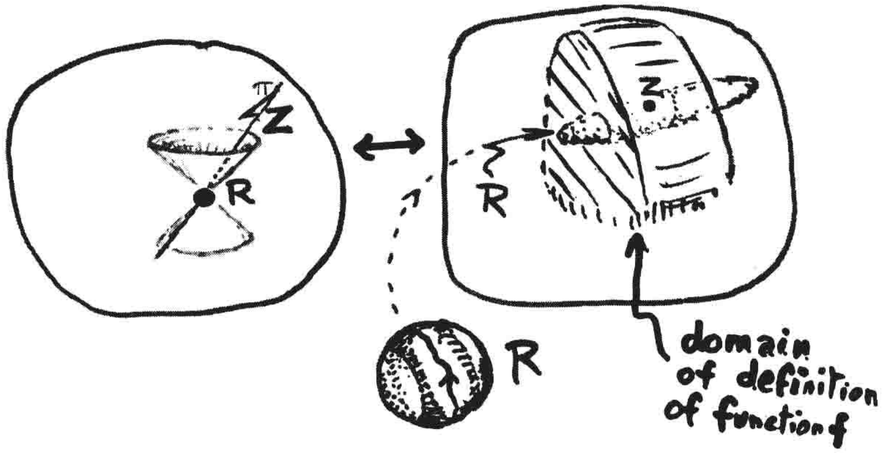


Fig.22
The geometry of the twistor contour integral

it intersects each Riemann sphere in $\mathbb{PT}$, representing a point of $\mathbb{M}$ (or $\mathbb{CM}$) where the field $\phi\ldots$ is to be defined, in an annular region.

More correctly f should be thought of as a representative of an element of (first) cohomology. The region U of $\mathbb{CM}$, in which $\phi\ldots$ is defined, corresponds to a region U of $\mathbb{PT}$, swept out by the lines (Riemann spheres) that represent the points of U. In the simplest situation, we cover U by a pair of open sets U_1 , U_2 (with $U_1 \cup U_2 = U$), and f is defined on the intersection $U_1 \cap U_2$. More general, we might have several open sets in the covering $\{U_i\}$, and then we require a whole collection of holomorphic functions $\{f_{ij}\}$, where $f_{ij} = f_{ji}$ defined on the intersection $U_i \cap U_j$ of a pair of the sets. See Fig. 23.

Twistor theory, the Einstein equations, and quantum mechanics. *Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation* (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, 1996

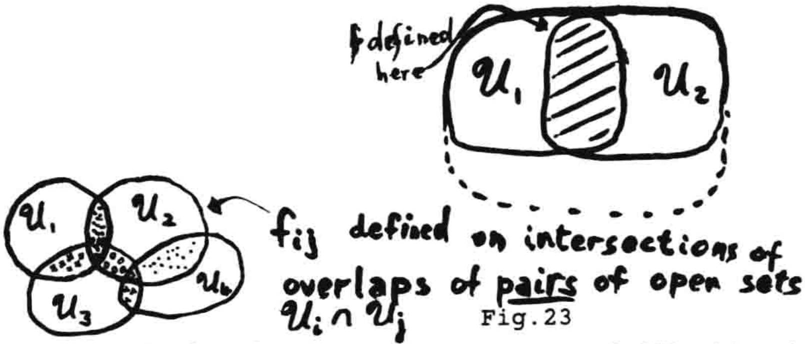

f defined here
$\mathcal{U}_1$
$\mathcal{U}_2$
$\mathcal{U}_1$
$\mathcal{U}_2$
$\mathcal{U}_3$
$\mathcal{U}_4$
f_{ij} defined on intersections of overlaps of pairs of open sets $\mathcal{U}_i \cap \mathcal{U}_j$

Fig.23

Strictly, f is a representative of (first) cohomology

The result of such a contour integral is always a solution of the massless free-field equations. These include the scalar wave equation (for spin zero), and I shall distinguish the positive helicity case from the negative helicity case by the use of primed rather than unprimed indices (where we assume that wavefunctions are given as complex fields of positive frequency). Thus we have (all field quantities being totally symmetric), for the massive free-field equations,

$$\nabla^{AA'}\phi_{AB\ldots L} = 0 \qquad (\text{helicity} < 0)$$

$$\Box\phi = 0 \qquad (\text{helicity} = 0)$$

$$\nabla^{AA'}\tilde{\phi}_{A'B'\ldots L'} = 0 \qquad (\text{helicity} > 0)$$

and we can distinguish the twistor contour integrals for negative helicity from those for positive helicity (zero helicity being a special case of each) by using the respective explicit forms

$$\text{helic.} \leqslant 0: \quad \phi_{AB\ldots L}(x) = \frac{1}{2\pi i}\oint \frac{\partial}{\partial\omega^A}\frac{\partial}{\partial\omega^B}\cdots\frac{\partial}{\partial\omega^L} f(\omega,\pi)\,\pi_{Q'}\,d\pi^{Q'}$$

$$\text{helic.} \geqslant 0: \quad \tilde{\phi}_{A'B'\ldots L'}(x) = \frac{1}{2\pi i}\oint \pi_{A'}\pi_{B'}\cdots\pi_{L'}\, f(\omega,\pi)\,\pi_{Q'}\,d\pi^{Q'}$$

where Z^α is taken incident with x^α

$$\omega^P = ix^{PP'}\pi_{P'}$$

The relationship between these two expressions is essentially consistent with the complex conjugation operation, since according to the twistor commutation rule $[Z^\alpha, \bar{Z}_\beta] = \hbar\,\delta^\alpha_\beta$ we represent $\bar{Z}_\alpha$ by $-\hbar\,\partial/\partial Z^\alpha$, so the complex conjugate of $\pi_{A'}$ is indeed represented by $\partial/\partial\omega^A$ (times the factor $-\hbar$), so that for full consistency a further constant factor may be included in the integrals, such as $(-\hbar)^{-n}$ in the first expression. Various familiar field equations, such as the Maxwell equations (for helicity ± 1) or the linearized Einstein equations (for helicity ± 2 written in the form of Bianchi identity applied to a weak-field curvature tensor K_{abcd}) are expressed by these massless free-field equations, where

$$F_{ab} = F_{AA'BB'} = \phi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\tilde{\phi}_{A'B'}$$

$$K_{abcd} = \phi_{ABCD}\,\varepsilon_{A'B'}\,\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\,\tilde{\phi}_{A'B'C'D'}.$$

For complex wave functions the tilde quantities are independent of the non-tilde quantities, but for classical real fields, the tilde is replaced by a bar — the symbol for complex conjugation.

We can now see how the "Mk.3" interpretation of a twistor as a spin-lovering operator is more precisely interpreted as *helicity raising*. We represent the twistor Y^α by the operator $Y^\alpha\partial/\partial Z^\alpha$ acting on twistor functions f(Z) corresponding to

$$\phi_{AB\ldots LM} \Longrightarrow \phi_{AB\ldots LM}\eta^M$$

or

$$\tilde{\phi}_{A'\ldots L'} \Longrightarrow \eta^M\nabla_{MM'}\phi_{A'\ldots L'} - i(n+1)\zeta_{(M'}\tilde{\phi}_{A'\ldots L')}$$

where $Y^\alpha = (\eta^A, \zeta_{A'})$. Corresponding, a *dual* twistor plays a role as a helicity *lowering operator*.

7. Positive and negative frequency splitting

Quantum (field) theory requires a notion of splitting field quantities into *positive* and *negative* frequency parts, to express quantum fields propagating into the future and past, respectively. In one dimension, we understand this splitting, using the notion of Fourier transform by interpreting positive frequency as superpositions of e^{-ipx} for $p > 0$, and negative frequency as superpositions of $e^{-i\tilde{p}x}$ for $\tilde{p} < 0$. For space-time fields in four-dimensions, the positive frequency fields are superpositions of $\exp(-ip_a x^a)$ where the momentum 4-vector p^0 is future pointing and negative frequency fields are superpositions of $\exp(-i\tilde{p}_a x^a)$ where $\tilde{p}^a$ is past-pointing (see Fig.24).

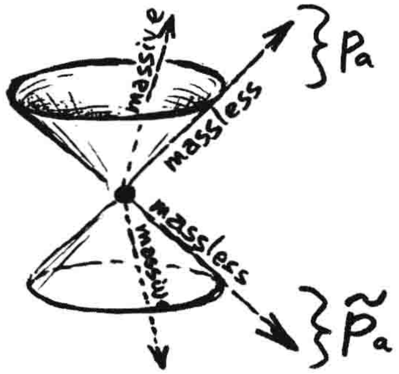


Fig.24
Momentum 4-vectors

There is also a more geometrical way of representing the notion of positive/negative frequency. In the 1-dimensional case, we can use the Riemann sphere - which, as we saw earlier (in Fig.10) can be related by stereographic projection, from the south pole, to the ordinary complex plane. In our previous picture, the equator on the sphere coincided with the unit circle on the plane. For convenience of description, however, it will be more appropriate to rotate the sphere so that the real axis (together with ∞) is represented as the equator (see Fig.25). The result that we shall use is that if a complex-valued function of a real variable is viewed as being defined on the equator of this Riemann sphere, then it has positive or negative frequency according to whether it can be extended holomorphically into the upper or lower hemispheres of the Riemann sphere.

Twistor theory provides us with a way of doing something remarkably similar, but now globally, for the whole of space-time, in the case of a massless field ϕ....(wave function) and its corresponding twistor function f.

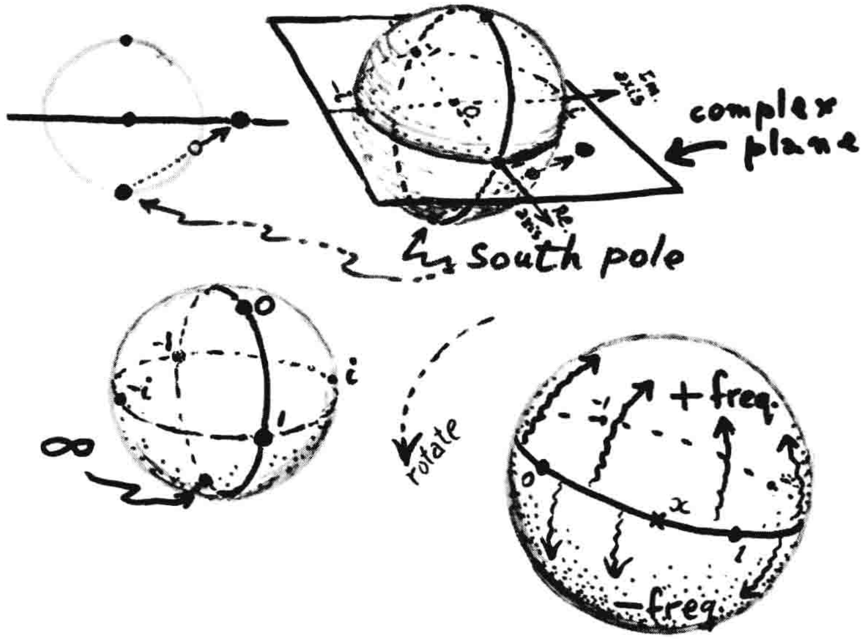


Fig.25

A complex function f(x) of a real variable x, defined on the equator of the Riemann sphere, is positive (negative) frequency if it extend as a complex holomorphic function for the upper (lower) hemisphere.

How do we ensure that f defines a wavefunction of positive or negative frequency? The answer is, regarding f as an element of first cohomology, that this cohomology element extends holomorphically from $\mathbb{PN}$ into the entire top-half [bottom-half] of twistor space ($\mathbb{PT}^+$ [or $\mathbb{PT}^-$]) if the wavefunction is of positive [negative] frequency. See Fig.26. This fact forms the basis of a good deal of the twistor approach to quantum field theory. Incomimg states, taken to have positive frequency, are represented as first cohomology elements on $\mathbb{PT}^+$ and outgoing states, taken to have negative frequency, as cohomology elements on $\mathbb{PT}^-$.

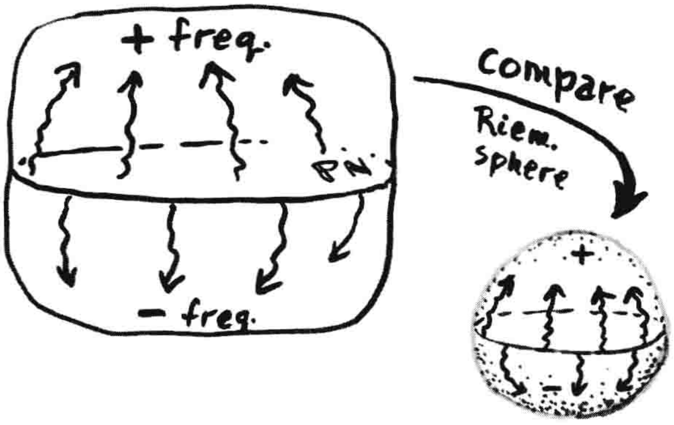


Fig.26

Positive-frequency twistor wavefunctions considered as elements of first cohomology, extend into the top half of twistor space if they have positive frequency and into the bottom half if negative frequency - similarly to the behaviour of ordinary functions of the Riemann sphere

Most of the detailed work on the twistor approach to quantum field theory has been following a perturbative approach using the twistor analogues of Feynman diagrams known as twistor diagrams. The subject hes been developed particularly by Andrew Hodges and his students, and some novel procedures for regularizing infinities have been discovered which have no direct analogues in conventional space-time, or momentum-space, descriptions. So far, the general approach lacks internally motivated driving principles and relies, instead, on translating directly from the conventional Feynman diagrams. An idea for possibly supplying such a driving principle was introduced by Michael Singer, and developed subsequently by Hodges and myself. In this approach, the (punctured) Riemann surfaces of conformal field theory are replaced by 3-complex-dimensional analogues obtained by glueing bits of projective twistor space together (see Fig.27).

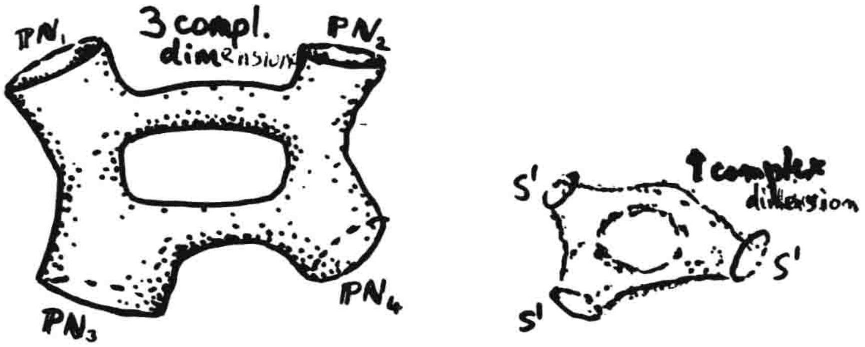


Fig.27
Generalization of conformal field theory for 4-dimensional space-time, where generalized (projective) twistor spaces take the place of Riemann surfaces

8. General relativity

There are numerous special applications of twistor theory to general relativity, but most of these will not be discussed here. Instead I shall try to indicate something of the general programme of bringing Einstein's general relativity into the scope of twistor theory. A major ingredient of this programme the "non-linear graviton construction" that I introduce some twenty years ago. This can be described roughly as follows.

Recall, first, the way in which (compactified) complex Minkowski space $\mathbb{CM}$ arises from $\mathbb{CP}^3$ (projective twistor space). Lines in $\mathbb{CP}^3$ represent points in $\mathbb{CM}$, where meeting lines represent null-separated points see (Fig.28). Now, let us try to represent a curved (complex) space-time by deforming the complex structure of $\mathbb{CP}^3$. In fact, by general theorems, there is no non-trivial complex deformation of $\mathbb{CP}^3$. However, there are non-trivial deformations of portions of $\mathbb{CP}^3$. We may consider a portion P of $\mathbb{CP}^3$ which is a tabular neighbourhood of a line in $\mathbb{CP}^3$ (representing an open neighbourhood of a point in $\mathbb{CM}$). We can deform P by cutting it (say) two (overlapping) pieces which are then re-glued, with the pieces somewhat displaced from their original attachment.

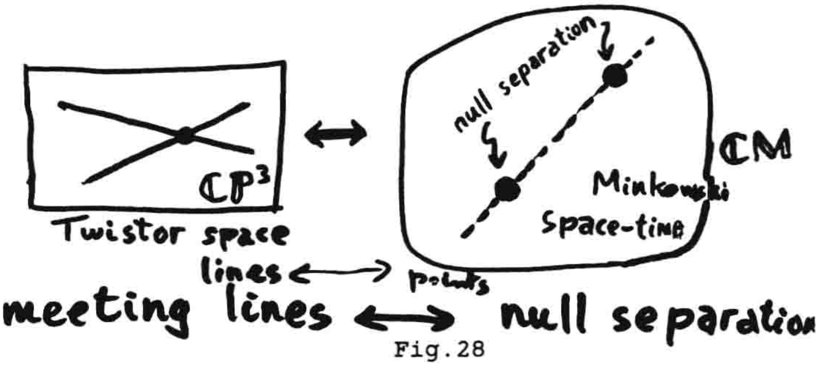

null separation
$\mathbb{CP}^3$
Twistor space
$\mathbb{CM}$
Minkowski Space-time
lines ⟷ points
meeting lines ⟷ null separation

Fig.28

Twistor representation of (complex) flat space-time

It follows from general theorems by Kodaira (and Spencer) that, provided the deformation is within some limit, there is a 4-parameter family of holomorphic curves in this re-glued space (which I shall call $\mathbb{P}\mathfrak{J}$) - distinguished merely by the fact that they are holomorphic and belong to the right topological class - which represents the points of a 4-dimensional complex manifold $\mathcal{M}$. The manifold $\mathcal{M}$ obtains a complex conformal metric, where null separated points in $\mathcal{M}$ are represented by meeting lines in $\mathbb{P}\mathfrak{J}$, just as was the case for $\mathbb{CM}$ (see Fig.29)

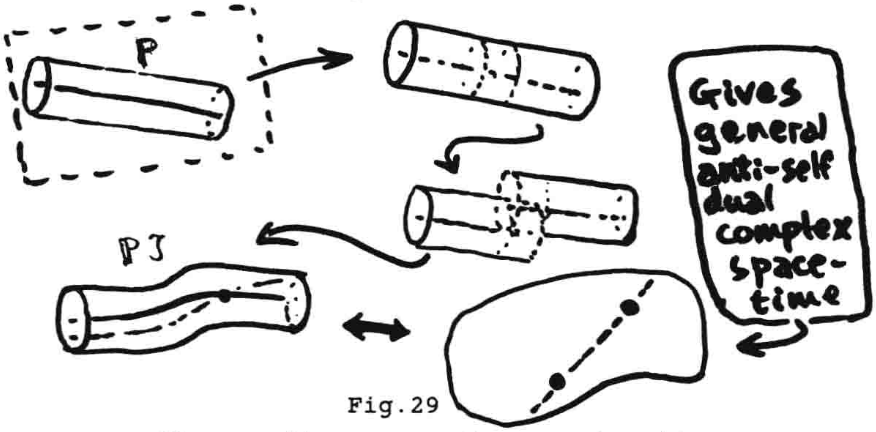

P
$\mathbb{P}\mathfrak{J}$
Gives general anti-self dual complex space-time

Fig.29

The non-linear graviton construction

In fact, this construction always provides a complex space-time $\mathcal{M}$ which has an (anti)-self-dual conformal curvature. We recall that the conformal curvature is described by the Weyl tensor $C^{a}{}_{bcd}$, which is the part of the Riemann tensor $R^{a}{}_{bcd}$ all of whose traces vanish. The symmetries and trace-free properties of C_{abcd} being the same as those of the weak-field vacuum curvature K_{abcd} considered earlier, there is a spinor description

$$C_{abcd} \rightsquigarrow \Psi_{ABCD}\,\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'},$$

where the symmetric spinors ψ_{ABCD} and $\tilde{\psi}_{A'B'C'D'}$ represent the anti-self-dual and self-dual parts of the conformal curvature, respectively. A complex space-time is conformally anti-self-dual if $\tilde{\psi}_{A'B'C'D'} = 0$ and conformally self-dual if $\psi_{ABCD} = 0$. It turns out that the integrability condition for the existence of a 3-parameter family of α-planes is $\tilde{\psi}_{A'B'C'D'} = 0$ and for the existence of a 3-parameter family of β-planes is $\psi_{ABCD} = 0$. In the non-linear graviton construction, there is a 2-parameter family of lines through each point of $\mathbb{P}\mathfrak{T}$. Thus each point Q of $\mathbb{P}\mathfrak{T}$ represents an entire 2-surface in $\mathcal{M}$, whose points are represented by the lines through Q. Since any pair of these lines intersect, it follows that any pair of points on the surface represented by Q must be totally null. Such a 2-surface is called an α-plane (or α-surface), and since $\mathbb{P}\mathfrak{T}$ is 3-dimensional, it follows that $\mathcal{M}$ contains a 3-parameter family of α-planes. Hence $\mathcal{M}$ is conformally anti-self-dual. Reversing the construction, we see that all conformally anti-self-dual complex space-times can be (locally) constructed in this way. See Fig.30.

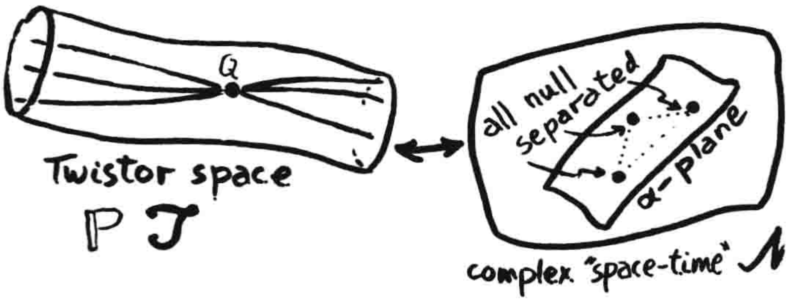


Fig.30
Each point of $\mathbb{P}\mathfrak{T}$ represents an α-plane in $\mathcal{M}$

The condition that $\mathcal{M}$ also satisfy the Einstein vacuum equations, i.e. be Ricci-flat, is most easily expressed in terms of the non-projective curved twistor space $\mathfrak{T}$. The space $\mathfrak{T}$ is, in effect, a complex line bundled over $\mathbb{P}\mathfrak{T}$. We can understand a point of $\mathfrak{T}$ as representing an α-plane in $\mathcal{M}$ together with a spinor $\pi_{A'}$ which is parallelly propagated over the α-plane, each tangent vector to the α-plane having the form $\lambda^A\pi^{A'}$, for some λ^A. The consistency of this notion follows from the fact that $\tilde{\psi}_{A'B'C'D'} = 0$, this providing the necessary integrability condition for parallel $\pi_{A'}$-spinors over α-planes. Now the condition that there is a parallelism for $\pi_{A'}$-spinors (locally) over $\mathcal{M}$ (not just over α-planes) is that $\mathcal{M}$ be both anti-self-dual and Ricci flat. For we have

$$[\nabla_{AA'}, \nabla_{BB'}]\pi_{C'} = 0$$

as the integrability condition for this. Splitting this into various parts, we see

$$\text{Part } [AB](A'B'C') = 0 \text{ by } \tilde{\Psi}_{A'B'C'D'} = 0;$$

$$\text{Part } [AB](A'B')^{B'} = 0 \text{ by } R = 0;$$

$$\text{Part } (AB)[A'B']C' = 0 \text{ by } R_{ab} - \tfrac{1}{4}g_{ab} = 0.$$

The remaining curvature ψ_{ABCD} represents the non-existence of β-planes in $\mathcal{M}$, i.e. the non-existence of holomorphic "planes" in $\mathbb{P}\mathfrak{T}$ (ruled by lines). We now have a notion of parallelism between α-planes (when their $\pi_{A'}$-spinors are parallel). This parallelism means that $\mathfrak{T}$ is fibred over the space of (non-zero) $\pi_{A'}$-spinors; equivalently, that $\mathbb{P}\mathfrak{T}$ is fibred over the Riemann sphere of projective $\pi_{A'}$-spinors (see Fig.31).

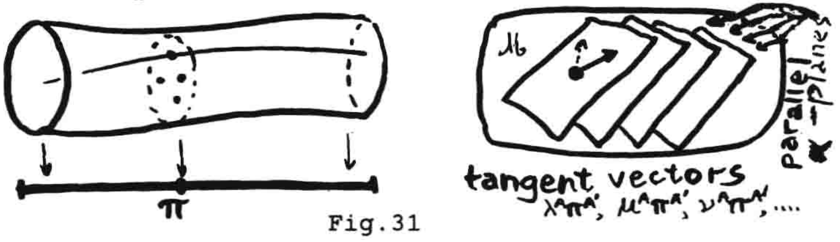


Fig.31

For $\mathcal{M}$ to be Ricci-flat the twistor space is fibred over the π-space, expressing a parallelism of α-planes in $\mathcal{M}$

The space $\mathfrak{J}$ has a local structure given by a simple 2-form τ, a 4-form Δ, and a vector field Y (the Euler operator whose integral curves provide the spaces $\mathbb{P}\mathfrak{J}$). This structure is locally the same as that of $\mathbb{T}$, where

$$\tau = (1/2)d\pi_{A'} \wedge d\pi^{A'}, \quad \Delta = (1/24)dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \epsilon_{\alpha\beta\gamma\delta}, \quad Y = Z^\alpha \partial/dZ^\alpha$$

The complex metric of $\mathcal{M}$ is determined by these forms. In the case of an infinitesimally deformed twistor space, there is a relation to a twistor function $f(z)$ degree + 2, the deformation being given by the glueing together of two portions of twistor space, the twistor Z^α of one space being matched to the twistor $\hat{Z}^\alpha$ of the other by

$$\hat{\omega}^A = \omega^A + \epsilon \varepsilon^{AB} \frac{\partial f}{\partial \omega^B}, \qquad \hat{\pi}_{A'} = \pi_{A'}$$

(see Fig.32). This provides an infinitesimal shunt in the twistor by $(\epsilon^{AB}\partial f/\partial\omega^B, 0)$, and it agrees with the formula

$$\psi_{ABCD} = \oint \frac{\partial}{\partial\omega} \cdots \frac{\partial}{\partial\omega} f \, \pi d\pi ,$$

described earlier, for the weak field (anti-self-dual) conformal curvature as helicity - 2 field.

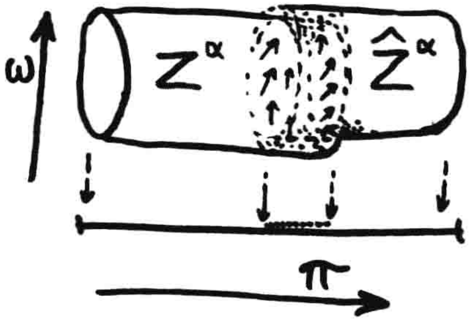


Fig.32

The weak field non-linear graviton.

Two patches of twistor space are displaced along a vector field given by $\epsilon^{AB}(\partial f/\partial\omega^A)\partial/\partial\omega^B$, where f has homogeneity + 2.

9. Ward construction for (anti-)self-dual Yang-Mills fields

Closely analogous to the non-linear graviton construction for anti-self-dual Einstein fields is Richard Ward's construction for anti-self-dual Yang-Mills fields. A Yang-Mills connection on an (analytic) 4-space $\mathcal{M}$ with anti-self-dual conformal curvature is anti-self-dual if and omly if it is integrable on α-planes (in its complexification $\mathbb{C}\mathcal{M}$ if $\mathcal{M}$ is real - and therefore conformally flat). This integrability implies that the bundle over $(\mathbb{C})\mathcal{M}$, upon which this connection acts, can be carried over to the twistor space $\mathbb{P}\mathfrak{T}$ (or $\mathfrak{T}$). See Fig.33.

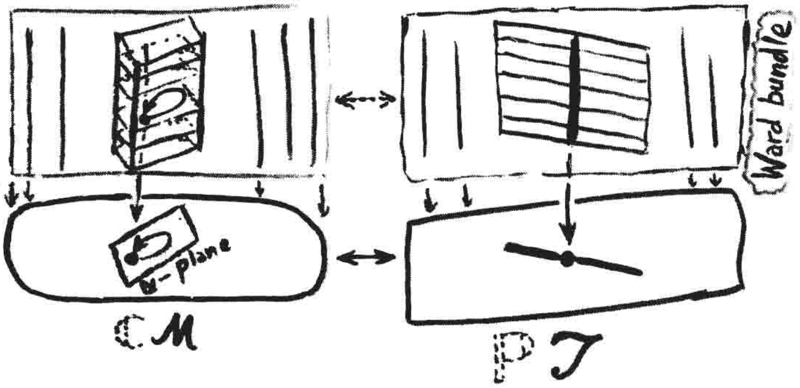


Fig.33

It turns out that the connection on the bundle over $(\mathbb{C})\mathcal{M}$ is completely determined by merely the holomorphic structure of the corresponding twistor bundle. moreover, this Yang-Mills connection is automatically anti-self-dual.

The Ward bundle can be constructed in terms of free transition functions and this fact leads to many applications to integrable systems (KdV, sine-Gordon non-linear Schrödinger, Toda lattice, Einstein with two commuting Killing vectors, etc.; work by Ward, Sparling, Mason, Woodhouse, and many others).

838 Twistor theory, the Einstein equations, and quantum mechanics. *Quantum gravity: Proc. Int. Sch. Cosmology and Gravitation* (1995). Eds A. Zichichi and T. Hans-Jurgen Sabbata, 228–75. Sci. Cult. Ser. Phys., 10, 1996

10. The general Einstein vacuum equations

The most promising idea for relating twistors to the Einstein (vacuum) equations $R_{ab} = 0$ is to continue together the following two facts.

(1) Twistors describe the *charges* for helicity 3/2 massless fields - in flat space $\mathbb{M}$

(2) The Einstein vacuum equations are *consistency conditions* for the equations describing massless helicity 3/2 fields in *curved* space-time $\mathcal{M}$.

Thus, we have a reasonably clear-cut (but difficult) programme for finding a full twistor description for the Einstein vacuum equations. This is illustrated in Fig.34. We first try to find the appropriate space of "charges" for massless fields in $\mathcal{M}$ (assumed Ricci-flat) of helicity 3/2. This space is our desired curved twistor space $\mathcal{T}$. We hope that $\mathcal{T}$ can somehow be specified in a natural way in terms of free functions. We then try to see how to reconstruct $\mathcal{M}$ from $\mathcal{T}$ in terms of some kind of global structure.

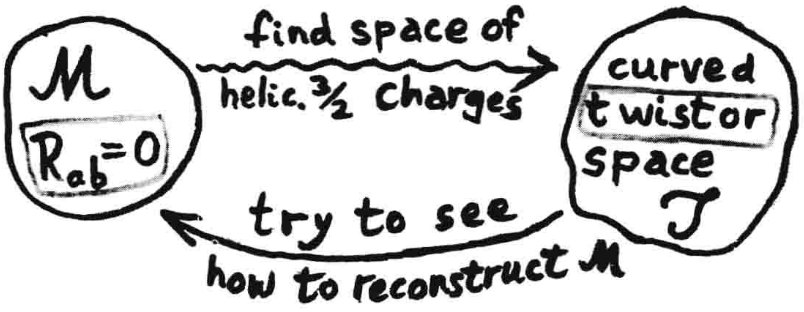


Fig.34

Programme for representing Ricci-flat space-times in terms of twistor space: highly non-local

This is still very much "work in progress", but it will be of some interest to consider, here, certain ideas that are involved. To understand what is meant by the term "charges", as used here, consider the case of spin 1. We imagine a region of space-time in which a free Maxwell field F_{ab} is defined, and which surrounds some world-tube containing sources to the field. We can obtain the (conserved) total charge of the sources by performing a Gauss integral

$$e = \oint {}^{*}F$$

(ignoring factors of 4π, etc.) over some suitable topological sphere S^2, within the region of definition of the Maxwell field, surrounding (i.e. linking) the sorce region. Here ${}^{*}F$ is the 2-form defined by the dual Maxwell field:

$$\underset{\sim}{{}^{*}F} = {}^{*}F_{ab}\, dx^a \wedge dx^b$$

$${}^{*}F_{ab} = \tfrac{1}{2} e_{abcd} F^{cd}$$

We can also see a self-dual field

$${}^{+}F_{ab} \left(= \tfrac{1}{2} F_{ab} - \tfrac{i}{2}\, {}^{*}F_{ab}\right) = \varepsilon_{AB}\, \widetilde{\varphi}_{A'B'},$$

which provides us with a complex charge whose real and imaginary parts are the magnetic charge μ and (minus) the electric charge e, where

$$\mu - ie = 2\oint {}^{+}\underset{\sim}{F} = \oint \underset{\sim}{\widetilde{\varphi}};$$

with

$${}^{+}F = {}^{+}F_{ab} dx^a \wedge dx^b \quad \text{and} \quad \tilde{\varphi} = 2\tilde{\varphi}_{A'B'} dx_A^{A'} \wedge dx^{AB'}$$

In flat space-time $\mathbb{M}$, we can do something similar for the charges for spin 2, to obtain the mass, momentum, and angular momentum for a linearized gravitational field, given by a (self-dual) spinor $\tilde{\psi}_{A'B'C'D'}$. We can use the same Gauss integral $\oint\tilde{\varphi}$ as before, where we spin-lower twice:

$$\widetilde{\varphi}_{A'B'} = \mu^{C'} \hat{\mu}^{D'} \widetilde{\psi}_{A'B'C'D'},$$

where μ and $\tilde{\mu}$ are solutions of the (complex conjugated) twistor equation for dual twistors $W_\alpha = (\lambda_A , \mu^{A'})$ and $\hat{W}_\alpha$. Similarly, we can obtain the charges for spin 3/2 by

$$\tilde{\varphi}_{A'B'} = \mu^{C'}\psi_{A'B'C'} .$$

Then

$$\oint \tilde{\varphi} = e(W_\alpha) = Z^\alpha W_\alpha ,$$

since e is linear in W_α . This defines the *twistor charge* Z^α .

Let us examine the equations for helicity 3/2 more carefully. The field is given by a symmetric spinor $\psi_{A'B'C'}$ satisfying

$$\nabla^{AA'}\psi_{A'B'C'} = 0 .$$

We need potentials for this field. There are two wersion of this: the Dirac and Rarita-Schwinger forms of the potentials. For simplicity (and other reasons of convenience) I shall use the Dirac form here. In the following, the ticked (✓) equations can be consistently used in Ricci-flat curved space-time whereas the ones without a tick must be discarded when $\mathcal{M}$ is not flat. We have a first potential

$$\sigma_{A'B'}{}^{C} = \sigma_{(A'B')}{}^{C} \qquad (\checkmark)$$

subject to

$$\nabla^{AA'}\sigma_{A'B'}{}^{C} = 0 \qquad (\checkmark)$$

$$\nabla_{CC'}\sigma_{A'B'}{}^{C} = \psi_{A'B'C'}$$

and a second potential

$$\rho_{A'}{}^{BC} = \rho_{A'}{}^{(BC)} \qquad (\checkmark)$$

subject to

$$\nabla^{AA'}\rho_{A'}{}^{BC} = 0$$

$$\nabla_{BB'}\rho_{A'}{}^{BC} = 2i\ \sigma_{A'B'}{}^{C} . \qquad (\checkmark)$$

There is a gauge freedom, given by

$$\sigma_{A'B'}{}^{C} \Longrightarrow \sigma_{A'B'}{}^{C}\ \nabla_{B'}^{C}\ \pi_{B'} \qquad (\checkmark)$$

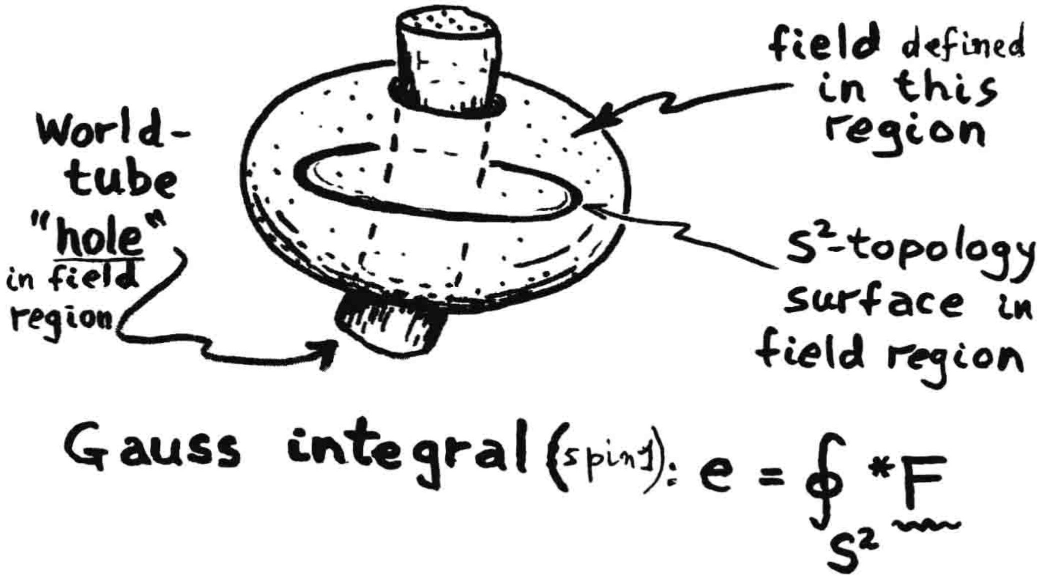

field defined in this region
World-tube "hole" in field region
S^2-topology surface in field region
Gauss integral (spin 1): $e = \oint_{S^2} {}^*\underset{\sim}{F}$

Fig.35

Charges: as properties of holes in regions of source free field where

$$\nabla^{BB'}\pi_{B'} = 0 \qquad (\checkmark)$$

and

$$\rho_{A'}^{\ BC} \Longrightarrow \rho_{A'}^{\ BC} - i\,\varepsilon^{BC}\pi_{A'} + \nabla_{A'}^{\ C}\,\omega^{B} \qquad (\checkmark)$$

where

$$i\,\nabla_{BA'}\omega^{B} = 2\pi_{A'}\ . \qquad (\checkmark)$$

The charges arise as a "gauge freedom of the second kind", namely the freedom in the gauge quantities which leaves the potentials unchanged. We can understand this in terms of the exact sequence

$$0 \longrightarrow \underset{\substack{\text{twistor}\\ \text{charge}}}{\left\{\begin{matrix}\Omega\\ \Pi\end{matrix}\right\}} \longrightarrow \underset{\substack{\text{gauge}\\ \text{quantities}}}{\left\{\begin{matrix}\omega\\ \Pi\end{matrix}\right\}} \longrightarrow \underset{\text{potentials}}{\left\{\begin{matrix}\rho\\ \sigma\end{matrix}\right\}} \longrightarrow \underset{\text{fields}}{\left\{\ \psi\ \right\}} \longrightarrow 0$$

The equations satisfied by the charges are

$$\nabla_{AA'}\Pi_{B'} = 0$$

$$\nabla_{AA'}\Omega^{B} = -\,i\,\varepsilon_{A}^{\ B}\,\Pi_{A'}\ ,$$

which are precisely the equations defining a twistor in $\mathbb{M}$.

Alternatively, we can use the Rarita-Schwinger form of the potentials. Here ρ, σ are not symmetric and the gauge is free. The equations for the potentials are a little more complicated than in the Dirac form, and I shall not use them here.

We cannot directly use the above exact sequence in curved (Ricci-flat) space-time because the inconsistency of the unticked equations. Nevertheless, it is useful to persue the topological implications of the existence of such a sequence. In $\mathbb{M}$, we can adopt a topological procedure for charges in flat space as "glitches" in the gauge quantities. See Fig.36. We take a region in $\mathbb{M}$ whose topology is that of a (tickered out) sphere, in which the field is globally defined. This is divided into two overlapping hemispherical regions in each of which there is a separate choice of potential for the field. The difference between the potentials in the two halves provides a zero-field potential on the annular overlap. This zero field potential is locally obtenaible from a pure gauge, but as one proceeds around the annulus this gauge jumps. The jump must be an element of the "gauge freedom of the second kind", i.e. an element of the charge space - a twistor.

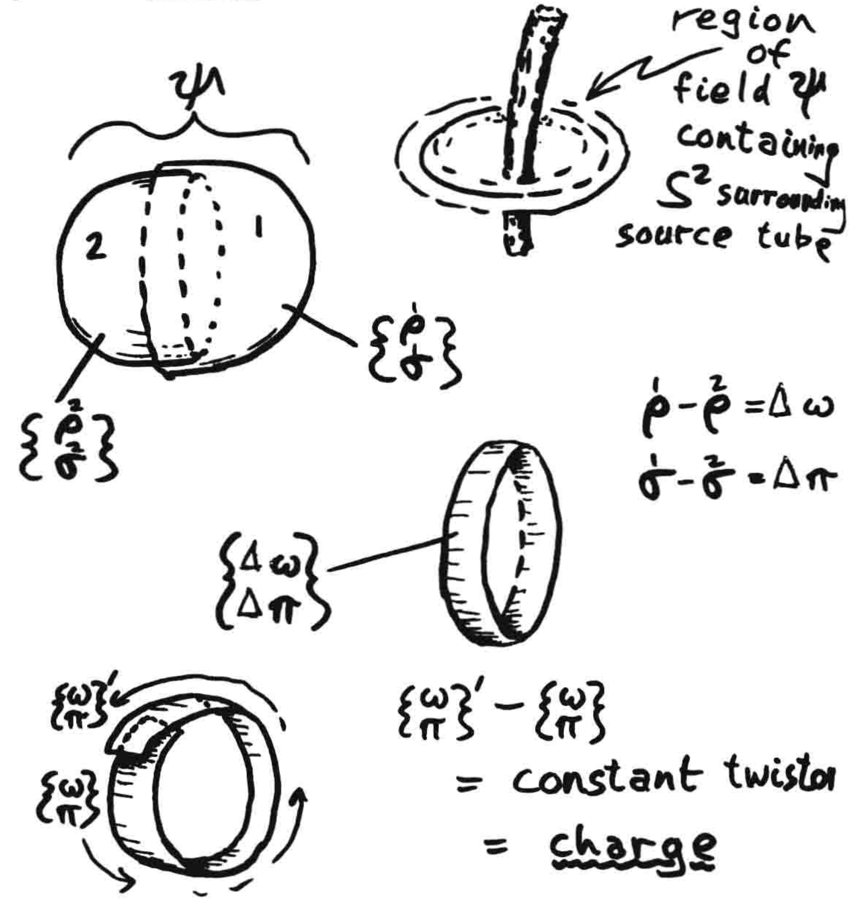

ψ
2
1
$\{\overset{1}{\rho} \ \overset{1}{\sigma}\}$
$\{\overset{2}{\rho} \ \overset{2}{\sigma}\}$
region of field ψ containing S^2 surrounding source tube
$\overset{1}{\rho} - \overset{2}{\rho} = \Delta\omega$
$\overset{1}{\sigma} - \overset{2}{\sigma} = \Delta\pi$
$\{\Delta\omega \ \Delta\pi\}$
$\{\omega \ \pi\}'$
$\{\omega \ \pi\}$
$\{\omega \ \pi\}' - \{\omega \ \pi\}$
= constant twistor
= charge

Fig.36

The procedure for obtaining the twistor charges for a helicicity 3/2 field in $\mathbb{M}$ as topological "glitches". Here, the notation $\Delta\omega + \nabla_{AA'}\omega^B + i\,\varepsilon_A^B\pi_{A'}$, $\Delta\pi = \nabla_{AA'}\pi_{B'}$ is being used.

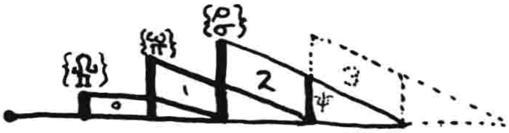
In curved (Ricci-flat) space-time $\mathcal{M}$, neither the fields ψ nor the constant twistors $(\Omega\ ,\ \Pi)$ exist. The best idea seems to be to go to null infinity $\mathscr{I}$ (assuming $\mathcal{M}$ is asymptotically flat), to find

"(Ω,Π)" at $\mathscr{I}$ and to define "ψ" merely as potential modulo gauge. Since, in this procedure, different (Ω,Π)s come about in different null directions and in different retarded times it is a non-linear process. However, the sequence just considered seems inadequate in that it does not "feel out" enough of the space-time. In $\mathcal{M}$, we cannot use "ψ" directly, but the final map of the sequence

$$0 \Longrightarrow \begin{Bmatrix}\Omega\\ \Pi\end{Bmatrix} \Longrightarrow \begin{Bmatrix}\omega\\ \pi\end{Bmatrix} \Longrightarrow \begin{Bmatrix}\rho\\ \sigma\end{Bmatrix}$$

still makes sense and we think of Ω and Π specified only at infinity. It appears that we need to lengthen this sequence in order to extend the "information" of (Ω , Π) back into the space-time. The above sequence only "feels out" 2-dimensional regions of space-time, whereas full initial data requires 3-dimensional regions.

We can redraw the above sequence (with its proposed extension) as



where the numerals denote the dimensions of the regions of $\mathcal{M}$ of relevance to the map in question. The key idea to finding the appropriate lengthening of this sequence seems to one due to Jorge Frauendiener (and George Sparling. The equation

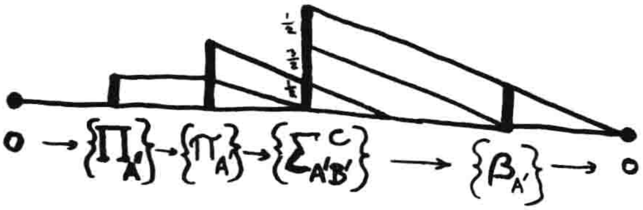
$$\nabla^{B'(B}\Sigma_{A'B'}{}^{C)} = 0 ,$$

where

$$\Sigma_{A'B'}{}^{C} = \Sigma_{B'A'}{}^{C} ,$$

is consistent (and conformally invariant) in arbitrary space-times.

In the Ricci-flat case, we have a (non-exact) sequence



(the fractions indicating the spins of the fields relevant to the indicated portions of the spaces). The gauge freedom is given by

$$\Sigma_{A'B'}{}^{C} \Longrightarrow \Sigma_{A'B'}{}^{C} + \nabla_{B'}{}^{C}\pi_{A'}$$

where

$$\nabla^{A'C}\pi_{A'} = 0 .$$

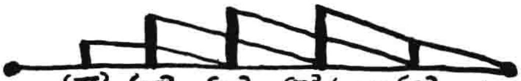
Also, we have

$$\nabla^{A'C}\beta_{A'} = 0$$

where

$$\beta_{A'} = \nabla^{B'}_{B}\Sigma_{A'B'}{}^{B} .$$

We can rewrite the above sequence as an exact sequence



$$\begin{tikzcd}
0 \arrow[r] & \{\Pi\} \arrow[r] & \{\pi\} \arrow[r] & \{\sigma\} \arrow[r] & \{\Sigma\}/\{\pi\} \arrow[r] & \{\beta\} \arrow[r] & 0
\end{tikzcd}$$

We can now introduce a second potential $P_{A'}^{BC}$ $(= P_{A'}^{(BC)})$, and we obtain a commutative diagram of exact sequences (which I shall not attempt to explain here in any detail):

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & \{\mathring{\Omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\rho}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{P}\}/\{\mathring{\omega}\} \arrow[r] \arrow[d, shift right, "i"'] & \{\mathring{\alpha}\} \arrow[r] \arrow[d, shift right, "i"'] & 0 \\
0 \arrow[r] & \begin{Bmatrix}\Omega\\ \Pi\end{Bmatrix} \arrow[r] \arrow[u, dashed, shift right, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\omega\\ \pi\end{Bmatrix} \arrow[r] \arrow[u, dashed, shift right, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\rho\\ \sigma\end{Bmatrix} \arrow[r] \arrow[u, dashed, shift right, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}P\\ \Sigma\end{Bmatrix}\Big/\begin{Bmatrix}\omega\\ \pi\end{Bmatrix} \arrow[r] \arrow[u, dashed, shift right, "\hat{q}"'] \arrow[d, shift right, "p"'] & \begin{Bmatrix}\alpha\\ \beta\end{Bmatrix} \arrow[r] \arrow[u, dashed, shift right, "\hat{q}"'] \arrow[d, shift right, "p"'] & 0 \\
0 \arrow[r] & \{\Pi\} \arrow[r] \arrow[u, dashed, shift right, "q"'] \arrow[d] & \{\pi\} \arrow[r] \arrow[u, dashed, shift right, "q"'] \arrow[d] & \{\sigma\} \arrow[r] \arrow[u, dashed, shift right, "q"'] \arrow[d] & \{\Sigma\}/\{\pi\} \arrow[r] \arrow[u, dashed, shift right, "q"'] \arrow[d] & \{\beta\} \arrow[r] \arrow[u, dashed, shift right, "q"'] \arrow[d] & 0 \\
 & 0 & 0 & 0 & 0 & 0 &
\end{tikzcd}$$

Here, i denotes injection and p, projection. The maps q and $\hat{q}$ derive from corresponding maps in $\mathbb{M}$,

$$\text{In } \mathbb{M}, \quad \begin{tikzcd} 0 \arrow[r] & \mathbb{S}^{A} \arrow[r, "i"] & \mathbb{T}^{\alpha} \arrow[r, "p"] \arrow[l, dashed, bend left, "\hat{q}"] & \mathbb{S}_{A'} \arrow[r] \arrow[l, dashed, bend left, "q"] & 0 \\ & \{\Omega\} & \left\{\begin{matrix}\Omega \\ \Pi\end{matrix}\right\} & \{\Pi\} & \end{tikzcd}$$

which represents points q, $\hat{q}$ in space-time, defined by

$$\Pi_{A'} \overset{q}{\longmapsto} (iq^{AA'}\Pi_{A'}, \Pi_{A'}), \quad (\Omega^{A}, \Pi_{A'}) \overset{\hat{q}}{\longmapsto} \Omega^{A} - iq^{AA'}\Pi_{A'}$$

The map q generalizes to anti-self-dual Ricci-flat #. This is in accordance with the way in which space-time points are represented in the non-linear graviton construction (see Fig.37): a point q of $\mathcal{M}$, in that construction, is a cross-section of ℑ, considered as a bundle over the (origin-free) primed spin space $\mathbf{S}_{A'} - \{0\}$ (π-space). This is a curved-space analogous of the map q considered above. This works when $\mathcal{M}$ is anti-self-dual. There should be a corresponding "googly" map $\hat{q}$ when $\mathcal{M}$ is self-dual. To see how space-time points are to be represented in twistors terms in the case of a general Ricci-flat space-time $\mathcal{M}$ is a major question for the future.

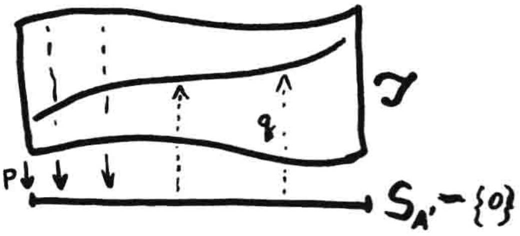


Fig.37

In the non-linear graviton, a space-time point is represented as a map q: a cross-section of ℑ as a bundle over π-space.

References

Aspect, A. , Grangier, P., and Roger, G. (1982). Experimental realization of Einstein-Podolsky-Rosen-Bohm *Gedankenexperiment*: a new violation of Bell's inequalities. *Phys Rev. Lett.* **48**, 91-4.

Bailey, T.N. and Baston, R.J., eds. (1990) *Twistors in Mathematics and Physics* London Mathematical Society Lecture Notes Series, 156 (Cambridge University Press)

Bell, J.S. (1966) On the problem of hidden variables in quantum theory, *Revs. Mod. Phys.* **38**, 447-52.

Buchdahl, H.A. (1958) On the compatibility of relativistic wave equations for particles of higher spin in the presence of a gravitational field *Nuovo Cim.* **10**, 96-103.

Deser, S. and Zumino, B. (1976) Consistent Supergravity *Phys. Lett.* ***62B***, 335-7.

Diósi, L. (1989) Models for universal reduction of macroscopic quantum fluctuations *Phys. Rev.* **A40**, 1165-74.

Dirac, P.A.M. (1936) Relativistic wave equations *Proc. Roy. Soc. London* **A155**, 447-59.

Eastwood M.G., Penrose, R., and Wells, R.O., Jr. (1981) Cohomology and massless fields, *Comm. Math. Phys.* **78**, 305-51.

Einstein, A., Podolsky, P. and Rosen, N. (1935) Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777-80.

Fierz, M. (1940). Uber den Drehimpuls von Teichlen mit Ruhemasse null und beliebigem Spin. *Helv. Phys. Acta* **13**, 45-60.

Ghirardi, G.C., Rimini, A. and Weber, T. (1986) Unified dynamics for microscopic and macroscopic systems. *Phys. Rev.*, **D34**,470.

Ghirardi, G.C., Grassi, R., and Rimini, A. (1990) Continuous-spontaneous-reduction model involving gravity. *Phys. Rev.*, **A42**,1057-64.

Hodges, A.P. (1982) Twistor diagrams, *Physica* **114A**, 157-75.

Hodges, A.P. (1985) A twistor approach to the regularization of divergences, *Proc. Roy. Soc. London* **A397**, 341-74. Mass eigenstatates in twistor theory, *ibid*, 375-96.

Hodges, A.P., Penrose, R., and Singer, M.A. (1989) A twistor conformal field theory for four space-time dimensions, *Phys. Lett.* **B216**, 48-52.

Huggett, S.A. (Ed.) (1995) *Twistor Theory* (Marcel Dekker, New York)

Huggett, S.A. and Tod, K.P. (1985) *An Introduction to Twistor Theory* London Math. Soc. student texts (L.M.S. publ.)

Hughston, L.P. (1979) *Twistors and Particles,* Lecture Notes in Physics No. 97 (Springer-Verlag, Berlin).

Hughston, L.P. and Ward R.S. (eds.) (1979) *Advances in Twistor Theory* (Pitman, London).

Julia, B. (1982) Système linéare associé aux équations d'Einstein *Comptes. Rendus. Acad. Sci. Paris, Sér. II,* 113-6.

Károlyházy, F. (1974) Gravitation and quantum mechanics of macroscopic bodies, *Magyar Fizikai Polyoirat* 12, 24.

Károlyházy, F., Frenkel, A. and Lukács, B. (1986) On the possible role of gravity on the reduction of the wave function in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Kibble, T.W.B. (1981) Is a semi-classical theory of gravity viable? In *Quantum Gravity 2: a Second Oxford Symposium*; eds. C.J. Isham, R. Penrose, and D.W. Sciama (Oxford Univ. Press, Oxford).

Kodaira, K. (1962) A theorem of completeness of characteristic systems for analytic submanifolds of a complex manifold. *Ann. Math.* (2) 75, 146-162.

Kodaira, K. and Spencer, D.C. (1958) On deformations of complex analytic structures I, II, *Ann. Math.* 67, 328-401, 403-466.

Majorana, E. (1932) Atomi orientati in campo magnetico variabile. *Nuovo Cimento* 9, 43-50.

Mason, L.J. and Hughston, L.P. (Eds.) (1990) *Further advances in twistor theory; Vol. I: The Penrose transform and its applications.* Pitman Research notes in Mathematics Series 231 (Longman Sci. & Tech., Harlow, Essex).

Mason, L.J., Hughston, L.P., and Kobak, P.Z. (Eds.) (1995) *Further advances in twistor theory; Vol. II: Integrable systems, conformal geometry and gravitation.* Pitman Research notes in Mathematics Series 232 (Longman Sci. & Tech., Harlow, Essex).

Mason, L.J. and Woodhouse, N.M.J. (1996) *Twistors, Self-Duality, and Integrable Systems* [approx. title] (Oxford Univ. Press).

Pearle, P. (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. Phys. Rev. A39, 2277-89.

Pearle, P. and Squires, E.J. (1995) Gravitaty, energy conservation and parameter values in collapse models. *Durham University preprint.* DTP/95/13.

Penrose, R. (1969) Solutions of the zero rest-mass equations, *J. Math. Phys.* 10, 38-9.

Penrose, R. (1976) Non-Linear gravitons and curved twistor theory *Gen. Rel. Grav.* 7, 31-52.

Penrose, R. (1982) Quasi-local mass and angular momentum in general relativity, *Proc. Roy. Soc. London* **A381**, 53-63.
Penrose, R. (1992) Twistors as spin 3/2 charges, in *Gravitation and Modern Cosmology* (P.G.Bergmann's 75th birthday vol.) eds. A. Zichichi, N. de Sabbata, and N.Sánchez (Plenum Press, New York).
Penrose, R. (1993) Gravity and quantum mechanics, in *General Relativity and Gravitation 1992. (Proc. 13th Int. Conf. on Gen. Rel. and Grav., Cordoba, Argentina. Part 1: Plenary Lectures.)* Eds. R.J.Gleiser, C.N.Kozameh, and O.M.Moreschi (Inst. of Phys. Bristol).
Penrose, R. (1994) *Shadows of the Mind; An approach to the missing science of consciousness* (Oxford Univ. Press, Oxford).
Penrose, R. (1996) On gravity's role in quantum state reduction *Gen. Rel. Grav.* (to appear).
Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge).
Penrose, R. and Rindler, W. (1986) *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge.
Percival, I.C. (1995) Quantum spacetime fluctuations and primary state diffusion. *Proc. R. Soc. Lond.* **A451**, 503-13.
Ward, R.S. (1977) On self-dual gauge fields, *Phys. Lett.* **61A**, 81-2.
Ward, R.S. and Wells, R.O. Jr (1989) *Twistor Geometry and Field Theory* (Cambridge University Press, Cambridge).
Woodhouse, N.M.J. and Mason, L.J. (1988) The Geroch group and non-Hausdorff twistor spaces *Nonlinearity* **1**, 73-114.
Zeilinger, A., Gaehler,R., Shull, C.G., and Mampe, W. (1988) Single and double slit diffraction of neutrons. *Revs. Mod. Phys.* **60**, 1067.

Document generated by Anna’s Archive around 2023-2024 as part of the DuXiu collection (https://annas-blog.org/duxiu-exclusive.html).

Images have been losslessly embedded. Information about the original file can be found in PDF attachments. Some stats (more in the PDF attachments):

```
{
  "filename": "NDA3ODg1Mzcuemlw",
  "filename_decoded": "40788537.zip",
  "filesize": 164993634,
  "md5": "86738b5d91672c48bd467982a4366d2e",
  "header_md5": "003bf3fd8ec7eb0e7724c39fd5afde6d",
  "sha1": "1b1101fba5d2c3f20cfe0fe6126f188055671158",
  "sha256": "4ccfdb127f6132cf53b982d0ccc2f2fb4aab28939dbd280ff1bb74a63bd8c7fe",
  "crc32": 2133447451,
  "zip_password": "",
  "uncompressed_size": 202648939,
  "pdg_dir_name": "ROGER PENROSE COLLECTED WORKS VOLUME 5 1990-1996_40788537",
  "pdg_main_pages_found": 849,
  "pdg_main_pages_max": 849,
  "total_pages": 874,
  "total_pixels": 5371848720,
  "pdf_generation_missing_pages": false
}
```